

# MIDRAS-selvitys

---

*Toimintamalli, etäkäyttöjärjestelmä ja niiden toteutus*

© CSC – Tieteen tietotekniikan keskus Oy ja Rekisteritutkimuksen tukikeskus (ReTki)

22.7.2011

## Sisällysluettelo

Sisällysluettelo .....	3
1. Johdanto.....	5
1.1. MIDRAS-selvityshanke .....	5
1.2. Selvityksen sisältö ja rakenne.....	5
2. MIDRAS-toimintamallin ja -järjestelmän tavoitteet.....	6
3. Kansainvälinen selvitys.....	7
3.1. Etäkäyttö kansainvälisenä muutossuuntana.....	8
3.2. Etäkäyttöjärjestelmien yhteiset piirteet.....	9
3.3. Eri maiden ratkaisut .....	9
3.4. Yhteenveto eri maiden ratkaisuista .....	13
4. Hankkeen aikana opittuja asioita .....	14
4.1. Tutkijoiden näkökulmia .....	14
4.2. Viranomaisten näkökulmia.....	17
4.3. Lakitekniset kysymykset .....	19
5. Selvityshankkeen aikana toteutettu pilotti ja siitä saatu palaute .....	20
5.1. Pilotin suunnittelu ja tarkoitus.....	20
5.2. Pilotin rajaukset ja toteutus .....	20
5.3. Pilottijärjestelmä ja sen käyttö.....	21
5.4. Tietotekninen ympäristö .....	22
5.5. Pilottijärjestelmän käyttäjäpalaute.....	23
5.6. Viranomaisten edustajien kokemuksia pilotista .....	31
6. Käyttäjä- ja järjestelmävaatimukset ja suunnittelua ohjaavat periaatteet .....	32
7. Suositeltu MIDRAS-toimintamalli ja etäkäyttöjärjestelmä.....	33
7.1. MIDRAS-järjestelmän yleiskuvaus.....	33
7.2. MIDRAS-järjestelmän palvelut .....	35
7.3. Toimintamallin prosessit ja toimijoiden vastuut .....	41
7.4. MIDRAS-järjestelmän tekninen toteutus .....	55
8. Etenemissuunnitelma.....	60
8.1. Toteutusvaiheet ja aikataulu .....	60
8.2. Sidosryhmät ja yhteistyöhankkeet.....	64
Lähteet .....	69
Liitteet .....	72

Liite 1. Sanasto .....	72
Liite 2. Vaihtoehtoiset mallit ja ratkaisut .....	75
Liite 3. Keskeiset tutkimuksessa käytettävät rekisterit ja rekistereihin perustuvat aineistot.....	86
Liite 4. Henkilötietojen luovutusta säätelevät lait .....	89
Liite 5. Organisaation ja MIDRAS-ylläpidon välinen esimerkkisopimus.....	91
Liite 6. Metatiedon vaatimusmäärittely.....	93
Liite 7. MIDRAS-metatiedon vastaavuus DDI3-elementteihin .....	99

## Taulukot

Taulukko 1 MIDRAS-toimintamallin suunnittelua ohjanneet tavoitteet.....	6
Taulukko 2 Etäkäyttöjärjestelmien yhteiset periaatteet .....	9
Taulukko 3 Eri maiden etäkäyttöjärjestelmien ratkaisut.....	13
Taulukko 4 Pilotin palvelimet, palvelut ja asennukset .....	23
Taulukko 5 Yksityskohtaisempaa käyttäjäpalautetta .....	26
Taulukko 6 Kooste etäkäyttöjärjestelmän suunnittelua ohjaavista periaatteista .....	32
Taulukko 7 Toimijoiden vastuut aineistojen kuvailussa .....	44
Taulukko 8 Toimijoiden vastuut lupamenettelyissä .....	46
Taulukko 9 Yhteenveto vastuista .....	47
Taulukko 10 Yhteenvetotaulukko toimijoiden vastuista aineistojen pseudonymisoinnissa ja yhdistelyssä ...	50
Taulukko 11 Toimijoiden vastuut aineistojen käsittelyssä .....	54
Taulukko 12 Eri pseudonymisointivaihtoehtojen vertailu .....	76
Taulukko 13 Pseudonymisointivaihtoehtojen vertailu .....	77
Taulukko 14 Aineistotoimitusvaihtoehtojen vertailu .....	79
Taulukko 15 Vahvan tunnistuksen vaihtoehtojen vertailu.....	81
Taulukko 16 Tutkimusympäristövaihtoehtojen vertailu .....	82
Taulukko 17 Tarkastellut etäpöytätekniikat .....	82
Taulukko 18 Aineistomuotovaihtoehtojen vertailu .....	83
Taulukko 19 Rajapintavaihtoehtojen vertailu .....	84
Taulukko 20 Keskeiset tutkimukseen käytetyt rekisterit.....	86
Taulukko 21 Henkilötietojen luovutusta säätelevät lait .....	89
Taulukko 22 metatietotaulukoiden termien merkitys.....	93
Taulukko 23 metatietoa kuvailevat tiedot.....	93
Taulukko 24 aineistoa kokonaisuutena kuvailevat tiedot .....	94
Taulukko 25 aineistossa olevaa muuttujaa kuvailevat tiedot .....	96
Taulukko 26 muuttujan koodaukseen käytettyä koodia kuvailevat tiedot .....	97
Taulukko 27 vastaavuustaulukossa käytettyjen merkintöjen selitys.....	99
Taulukko 28 MIDRAS- ja DDI3-elementtien vastaavuus.....	99

# 1. Johdanto

## 1.1. MIDRAS-selvityshanke

MIDRAS-selvityshankkeen (Micro Data Remote Access System) tavoitteena on ollut suunnitella toimintamallia ja pilotoida etäkäyttöjärjestelmää, jonka avulla viranomaisrekistereihin kerättyjä luvanvaraisia tietoja voidaan käyttää tutkimustarkoituksiin nopeasti, helposti ja tietoturvallisesti, yksilöiden tietosuoja varmistuen. Opetusministeriön rahoittaman hankkeen ovat toteuttaneet CSC - Tieteen tietotekniikan keskus ja Rekisteritutkimuksen tukikeskus (ReTki). Hankkeelle asetetut tavoitteet olivat:

- Etäkäyttöjärjestelmän toimintamalliehdotuksen laatiminen
- Vastaavien järjestelmien ja eri toimintamallivaihtoehtojen selvittäminen ja esittely
- Selvitys eri toimijoiden mahdollisuuksista ja halukkuudesta osallistua järjestelmän rakentamiseen ja aineistojen toimittamiseen, sisältäen myös mahdollisten lainsäädännöllisten esteiden ja muutostarpeiden selvittämisen
- Pilottijärjestelmän rakentaminen.

Rekisteriaineistolla tarkoitetaan tässä raportissa viranomaisten hallussa olevia, julkisin varoin hallinnollisiin ja tilastotarkoituksiin rekistereihin kerättyjä yksikkötasoisia aineistoja, joiden käyttö on luvanvaraista. (Selvityksessä käytetty sanasto on esitelty tarkemmin liitteessä 1.)

Selvityksessä on hankkeelle asetettujen tavoitteiden mukaisesti keskitytty luvanvaraisten, yksikkötasoisien rekisteriaineistojen tutkimuskäytön edistämiseen. MIDRAS-järjestelmää voidaan kuitenkin hyödyntää myös muuntyyppisten aineistojen luovuttamisessa sekä yleiskäyttöisempänä tutkimus- ja analyysialustana esimerkiksi viranomaisten selvitystöissä.

Hankkeen ehdottama toimintamalli helpottaa ja monipuolistaa rekisteritutkimuksen tekemistä, tarjoaa tutkimuksen tueksi uuden teknisen ratkaisun ja parantaa julkisin varoin kerätyn tiedon saatavuutta ja uudelleenkäyttömahdollisuuksia. Toimintamalli parantaa erityisesti tietosuojajäleessä arkaluontoisten aineistojen käytettävyyttä ja yhdistämistä useammilta viranomaisilta.

Hankkeen luoman vision mukaan: *Kansallisen etäkäyttöjärjestelmän kautta viranomaisten hallussa olevat, luvanvaraiset, yksikkötasoiset aineistot ovat käytettävissä kattavasti, turvallisesti, kustannustehokkaasti ja helposti tutkimukseen.*

## 1.2. Selvityksen sisältö ja rakenne

MIDRAS-hankkeen tulosten esittely on jaettu kahteen raporttiin: MIDRAS-loppuraporttiin ja tähän MIDRAS-selvitykseen. MIDRAS-loppuraportissa kuvataan selvityksen tulokset yleisemmällä tasolla ja vedetään yh-

teen hankkeen antamat suositukset. Tässä selvityksessä esitellään puolestaan yksityiskohtaisemmin selvityshankkeessa ilmi tulleet asiat sekä niiden perusteella suositettava rekisteritietojen etäkäyttöjärjestelmän toimintamalli ja sen toteutussuunnitelma.

Selvitys on suunnattu erityisesti etäkäyttöjärjestelmän tuleville kehittäjille sekä muille järjestelmän kehityksestä kiinnostuneille. Loppuraportti on suunnattu kohdennetummin tietohallinnon organisoinnista päättävälle ja toiminnan rahoittajille.

Raportin rakenne on seuraava:

- Luku 2 kuvaa MIDRAS-toimintamallin ja etäkäyttöjärjestelmän suunnittelun pohjana olleet tavoitteet
- Luku 3 esittelee kansainvälisen selvityksen tulokset
- Luku 4 käy läpi selvityshankkeen aikana opitut asiat
- Luku 5 sisältää selvityshankkeen aikana toteutetun pilotin kuvauksen ja pilottijärjestelmän käyttäjäpalautteen
- Luku 6 esittelee selvityshankkeen perusteella laaditut käyttäjävaatimukset ja suunnitteluperiaatteet toimintamallin ja etäkäyttöjärjestelmän rakentamiselle
- Luku 7 määrittelee suositetun MIDRAS-toimintamallin, etäkäyttöjärjestelmän sekä palvelu-, prosessi- ja tekniset kuvaukset
- Luku 8 sisältää etenemissuunnitelman sekä keskeiset sidosryhmät, joiden kanssa toimintaa tulee jatkossa yhteistyössä edistää
- Liitteissä on esitelty selvityshankkeen aikana pohdinnassa olleet vaihtoehtoiset toteutusratkaisut, tutkimuskäytössä olevat rekisterit, rekistereiden käyttöä määrittävät lait, sopimusesimerkkejä ja metatietojen vaatimusmäärittely

## 2. MIDRAS-toimintamallin ja -järjestelmän tavoitteet

MIDRAS-toimintamallin keskeinen tavoite on se, että aineistojen saatavuus paranee, eri osapuolien työmäärä vähenee eikä mikään asia huononnu nykyhetkestä. MIDRAS-toimintamallia on hankkeen aikana suunniteltu taulukossa 1 lueteltujen tavoitteiden pohjalta. Taulukossa esitellään myös mahdollisia toteutustapoja haasteisiin vastaamiseksi.

**Taulukko 1 MIDRAS-toimintamallin suunnittelua ohjanneet tavoitteet**

Tavoite	Mahdollisia toteutustapoja
Rekisteriaineistojen saatavuuden parantaminen	Useiden viranomaisten tiedot voi saada yhden palvelun kautta. Toimintatapojen vakiointi esim. tutkimusluvuissa ja tietojen lähe-

	tyksessä ja yhdistelyssä nopeuttaa kommunikaatioprosesseja. Aineistojen kuvailuja kokoava aineistokatalogi lisää tietoisuutta käytettävissä olevista tutkimusaineistoista.
Rekisteritutkijoiden aineistojen laadun paraneminen	Ajantasaisemmat aineistot saadaan automatisoimalla tutkimusaineistojen tuotantoprosessia ja tukemalla tietojen hakemista suoraan lähderekistereistä. Tietotekniset aineistojen toimituskanavat helpottavat aineistojen korjausta ja päivitystä. Paremmen tietosuojan turvin voidaan tarjota tutkijoille laajempia aineistoja.
Arkaluonteisten aineistojen käsittelyn tietosuojan paraneminen	Tiedot siirretään kryptografisesti suojattuina. Analysointiprosessista tehdään läpinäkyvää aineistojen tuottajille. Tutkimusaineistot hävittämiskäytäntö vakioidaan. Aineistot lähetetään ilman henkilötunnuksia.
Rekisteritutkimuksen luotettavuuden, toistettavuuden ja säilytyksen paraneminen	Käytetyt menetelmät ja tulokset säilytetään keskitetysti.
Uuden ja uudenlaisen tutkimuksen rohkaisu	Aineistokatalogi kattaa erilaisia aineistoja, joiden yhdistely voi tuottaa uutta tietoa. Aiempien tutkimusten ja menetelmien arkisto tarjoaa resursseja tutkimukseen.
Tämänhetkistä parempi ympäristö aineistojen työstämiseen	Tilasto- ja analyysiohjelmit tarjotaan keskitetysti. Tarjotaan tehokkaat laskentaresurssit. Tutkimusryhmä pääsee eri paikoista samaan tutkimusympäristöön.
Järjestelmä, joka palvelee aineistontuottajien tutkimuspalveluita	Aineistojen luovuttamiseksi luodaan tietoturvallinen, keskitetty ratkaisu.
Järjestelmän laajennettavuus kattamaan uusia tarpeita ja aineistotyyppjä	Järjestelmäarkkitehtuuri suunnitellaan yleiskäyttöiseksi.

### 3. Kansainvälinen selvitys

Tässä luvussa esitellään eri maissa käytössä olevia rekisteritietojen etäkäyttäjärjestelmiä sekä näiden järjestelmien yhtäläisyyksiä, eroja, kustannus- ja rahoitusmalleja. Tarkasteluun on valittu Tanskan, Ruotsin, Hollannin, Englannin, Australian ja Kanadan etäkäyttäjärjestelmät. Lähteinä on käytetty projektin käyttöön saamia dokumentteja eri etäkäyttäjärjestelmistä ja niiden ratkaisuista (esim. Statistiska centralbyrån, 2003 ja 2011, Fomkin, 2009), projektin aikana käytyjä keskusteluja Ruotsin Mona-järjestelmän ja CODIR-projektin (Cross-Organizational Database Infrastructure for register-based Research) kehittäjien kanssa, International working group on Microdata Access-ryhmän tiedotteita (Australian Bureau of Statistics, 2010) sekä lehtiartikkeleita (Statistical Journal of IAOS, 2009).

### 3.1. Etäkäyttö kansainvälisenä muutossuuntana

Vuonna 2001 Tanskan tilastovirasto avasi ensimmäisenä tilastovirastona maailmassa tutkijoiden käytettäväksi tarkoitetun rekisteritietojen etäkäyttöjärjestelmän. Aikaisemmin tutkijat olivat käyttäneet Tanskan tilastoviraston rekisteriaineistoja tilastoviraston tutkimuslaboratoriossa. Se, että tutkimusta pystyi nyt tekemään omalta tietokoneelta, oli valtava edistysaskel. Etäkäyttöjärjestelmän avaamisen seurauksena tutkimusten määrä lisääntyi Tanskassa merkittävästi (Borchsenius, 2005). Ennen tätä varsinkin taloustutkijat eri maista olivat jo käyttäneet Luxembourg Income Studyn (2000) tarjoamaa etäajojärjestelmää (remote execution).

Tanskan onnistunut esimerkki kannusti etäkäyttöjärjestelmän käyttöönottoon myös muiden maiden tilastovirastoissa. Etäkäyttöjärjestelmä otettiin käyttöön esimerkiksi Englannissa 2003, Ruotsissa 2005 ja Hollannissa 2006. Tilastovirastoissa käynnistyi samalla myös ajattelutavan muutos. Aikaisemmin toiminnan suurimpana riskinä pidettiin liian tarkkojen yksilötasoisien aineistojen luovuttamista ja kansalaisten luottamuksen menettämistä tämän seurauksena. Uuden ajatustavan mukaan yhteiskunnan kannalta riskialttiimpaa on, jos aineistoja ei hyödynnetä tutkimuksiin tietosuojakysymysten takia. Riskiksi tunnistettiin myös se, että tutkijat käyttävät tutkimukseen aineistoja, joita on helpompi saada, mutta jotka eivät sovellu tutkimuksiin yhtä hyvin kuin tilastovirastojen keräämät tiedot. (Tam, 2009)

Voidaankin todeta, että riskien välttämiseksi on suunnattu kohti riskien hallintaa ja yksilö- tai yksikötasoisien aineistojen laajempaa käyttöä. OECD:n ja Australian tilastoviraston aloitteesta vuonna 2009 perustettiin International working group on Microdata Access -ryhmä. Ryhmän tavoitteena on kartoittaa eri maiden tilastovirastojen etäkäyttöön liittyvät järjestelmät ja tuoda esille uusia käytäntöjä ja teknologioita. Euroopan tilastovirastolla Eurostatilla on suunnitelmassa rakentaa infrastruktuuri salassa pidettävän tiedon etäkäyttöä varten (Reuter & Museux, 2009). Samalla ESFRI-projektissa 'Data without Boundaries' suunnitellaan toimintamallia Euroopan-laajuiselle hajautetulle etäkäyttöjärjestelmälle. Projekti aloitti toimintansa keväällä 2011 ja on 12 maan tilastovirastojen ja tietoarkistojen yhteisprojekti. (Dwbproject, 2011.) Myös EU:n vuoden 2009 tilastoasetuksen (N:o 223/2009) tavoitteena on tilastovirastoihin kerätyn yksikötasoisien aineiston käyttäminen laajemmin tutkimuksissa.

Nämä kehityssuuntaukset kuvaavat muutoksia, jotka ovat tapahtuneet eri maiden tilastovirastojen tietojen luovutuskäytännöissä. Muutokset ovat hankeselvityksen kannalta keskeisiä, sillä tilastovirastojen aineistot ovat pitkälti samanlaisia kuin ne aineistot, joita on tarkoitus käyttää MIDRAS-järjestelmän välityksellä: aineistot sisältävät joko henkilöiden tai yritysten yksikötasoisia salassa pidettäviä tietoja.

Terveystietojen, varsinkin biopankkeihin tallennettujen tietojen, tutkimiseen on lisäksi suunnitteilla ja käytössä erilaisia etäkäyttöjärjestelmiä tilastovirastojen ulkopuolella (BioGrid Australia, 2010). Vuonna 2006

Suomi osallistui seitsemän muun maan kanssa niin sanotun TwinNet -federaation perustamiseen, jossa luotiin järjestelmä eri maiden kaksosrekisterien tietojen käyttämiseksi yhdessä (esim. Johansen & Litton, 2005). Ruotsissa on parhaillaan menossa CODIR-projekti (Fomkin, 2009), jonka tarkoituksena on mahdollistaa eri paikoissa sijaitsevien aineistojen käyttäminen yhdessä. Näin tutkija pystyisi käyttämään saman etäkäyttäjärjestelmän avulla sekä Ruotsin tilastovirastossa sijaitsevaa aineistoa että terveystietoja Ruotsin sosiaali- ja terveystieteiden virastosta.

### 3.2. Etäkäyttäjärjestelmien yhteiset piirteet

Tarkastelun kohteena olevissa maissa on havaittavissa selkeästi samankaltaisia, etäkäyttäjärjestelmiä koskevia periaatteita.

**Taulukko 2 Etäkäyttäjärjestelmien yhteiset periaatteet**

<b>Etäkäyttäjärjestelmien yhteisiä periaatteita</b>
Aineistot sijaitsevat etäkäyttäjärjestelmän palvelimilla.
Tutkija ei voi kopioida omalle työasemalleen etätyöpöydällä käyttämiään aineistoja.
Tulokset tarkistetaan ennen tutkijalle luovuttamista (ainakin satunnaisesti).
Aineistot eivät sisällä suoria tunnuksia (henkilötunnuksia, yritystunnuksia).
Tutkijat työskentelevät etätyöpöydällä järjestelmän tarjoamilla ohjelmissa.
Ainoastaan hyväksytyt/käyttöluvan saaneet tutkijat (tai viranomaisten edustajat) pääsevät käyttämään aineistoja.
Käyttäjät tunnistetaan vahvalla tunnistautumisella.
Tutkijat allekirjoittavat salassapitositoumuksen ja heille annetaan tietosuojaan liittyvää ohjausta.
Tutkijoiden käyttämät aineistot on eroteltu muista tilastoviraston aineistoista.
Järjestelmän kaikki tietoliikenneyhteydet ovat suojattuja ja salakirjoitettuja.

### 3.3. Eri maiden ratkaisut

Vaikka perusratkaisut ovatkin samankaltaisia, on jokainen maa tehnyt omia ratkaisujaan lainsäädäntönsä, rahoitustilanteen ja käytettävissä olevien aineistojen mukaan. Seuraavassa tarkastellaan eri etäkäyttäjärjestelmien käyttäjiin, tarjolla oleviin aineistoihin, tietosuojaan, tekniseen järjestelmään ja rahoitukseen liittyviä ratkaisuja.

#### 3.3.1. Käyttäjät

Useimmissa maissa etäkäyttäjärjestelmien tarkoituksena on ollut palvella niin tutkijoita kuin viranomaiskäyttäjiä. Ensisijaisesti etäkäyttäjärjestelmät on kuitenkin tarkoitettu tutkijoiden käytettäväksi.

Tanskassa käyttäjäksi hyväksytään ainoastaan hyväksytyt tutkimuslaitoksen tai viraston tutkijat. Australiassa tehdään laitoskohtaisia aineistoluovutuksia. Hollannissa järjestelmä on tutkijoiden lisäksi myös tiettyjen

viranomaisten käytössä (Hoeve, 2009). Hollannin tilastovirastossa on kuitenkin esitetty huoli siitä, että viranomaiskäytössä etäkäyttöjärjestelmän välityksellä saadut tiedot vääristetään tukemaan omaa poliittista kantaa. Tästä syystä Hollannissa on erillinen luparyhmä, joka tarkistaa viranomaisilta tulleet lupahakemukset. Lisäksi kaikilta etäkäyttöjärjestelmän käyttäjiltä vaaditaan tutkimustulosten julkistamista.

Ruotsin Mona-järjestelmän käyttäjistä suurin osa on tutkijoita. Neljännes käyttäjistä tulee kuitenkin valtiollahinnosta, erityisesti kunnista, joissa laaditaan Mona-järjestelmän välityksellä oman alueen tilastotaulukoita käytettävissä olevista aineistoista. Tilastotaulukoiden laadinnassa kunnat käyttävät Mona-järjestelmän tarjoamaa Australian tilastoviraston suunnittelemaa SuperCross-taulukointiohjelmaa. Australian tilastovirastossa on todettu, että suurin osa etäkäyttöjärjestelmän käyttäjien lopputuoksista on tilastotaulukoita. (Tam, 2009; Farley-Larmour&Gare, 2009).

Englannissa kehitettiin oma etäkäyttöjärjestelmä Virtual Microdata Laboratory (VML) lähtökohtaisesti viranomaiskäyttöä varten (Office for National statistics, 2008; Richie, 2009). Englannin tilastoviranomaisten lisäksi myös muut valtion virastot ovat liittäneet aineistojaan järjestelmään. Tutkijat ovat päässeet käyttämään aineistoja VML-järjestelmän välityksellä ainoastaan eri virastoissa sijaitsevilta yhteyskoneilta. Tähän tilanteeseen on tulossa muutos, sillä Englannin tietokonearkisto on avannut etäkäyttöjärjestelmän (Secure Data Service, SDS) (UK Data Archive, 2011). SDS on kopio VML:stä, jonka välityksellä tutkijat pääsevät omilta tietokoneiltaan käyttämään joitain samoja aineistoja kuin mitä VML:ssä on mahdollista käyttää.

Etäkäyttöjärjestelmien laajempi käyttöönnotto on nähty eri maissa kuitenkin ongelmallisena ulkomaisille tutkijoille. Esimerkiksi Tanskassa ja Ruotsissa tutkija ei saa käyttöoikeutta etäkäyttöjärjestelmässä oleviin tietoihin, mikäli hän ei ole tanskalaisen tai ruotsalaisen tutkimuslaitoksen tutkija... Ruotsissa on mahdollista vielä saada aineistoja perinteisesti CD:llä, mutta Tanskassa etäkäyttö on ainoa tapa käyttää rekisteriaineistoja. Tämän takia Tanska on jouduttu jättämään joidenkin yhteispohjoismaisten hankkeiden ulkopuolelle. Hollannissa ei ole rajattu käyttäjiä näin tiukasti: yksi Hollannin etäkäyttöjärjestelmän käyttäjistä onkin italialainen tutkimuslaitos. Australian tilastovirasto, joka ei lähtökohtaisesti salli tietojen luovutusta ulkomaille, on tehnyt sopimuksen Uuden-Seelannin tilastoviraston kanssa siitä, että kummankin maan tutkijat pääsevät käyttämään etäkäyttöjärjestelmän avulla toisen maan aineistoja.

### 3.3.2. Aineistot

Etäkäyttöjärjestelmien välityksellä tarjotaan tavallisimmin ns. valmisaineistoja, jotka on kehitetty tilastovirastojen tilastointia varten keräämistä aineistoista. Nämä aineistot on muodostettu joko hallinnollisista rekistereistä, kuten yleensä Pohjoismaissa, tai laajoista kyselyaineistoista. Esimerkiksi Hollannissa on käytettävänä yli 800 valmisaineistoa ja Australian tilastovirastossa on noin 120 CURF (Confidentialised Unit Record Files) -tiedostoa.

Tanskan tilastovirasto on muodostanut 17 eri aihealueesta yhteensä 250 valmista tutkimusaineistoa, joista poimitaan tutkijalle tutkimuksessa tarvittavat muuttujat tutkijan tarpeen mukaan (need to know –periaate). Nämä tutkijoille muodostetut tutkimusaineistot säilötään tilastoviraston palvelimille, jotta ne voidaan tarjota muiden tutkijoiden käytettäviksi kahden vuoden päästä siitä, kun aineisto on alun perin muodostettu. Ruotsissa tutkijalle avataan näkymä valmisaineistossa oleviin muuttujiin sen mukaan, mitä tietoja tutkija on ilmoittanut tarvitsevänsä. Tämän jälkeen tutkija pystyy itse yhdistämään eri tauluissa olevat tiedot, sillä Mona-järjestelmän aineistovalikoimassa henkilöille on annettu pseudonyymit. Myös Hollannin järjestelmässä tutkija yhdistää pseudonyymeja käyttäen eri valmisaineistot toisiinsa.

Metatietojärjestelyt vaihtelevat maittain. Hollannissa valmisaineistoja koskevat kuvaukset kerätään aineistokatalogiin. Aineistojen kuvailua varten Hollannissa ja Australiassa on suunniteltu DDI3 ja SDMX -metatietomäärittelyiden käyttämistä. Ruotsissa ja Tanskassa tilastovirastot ovat kuvanneet verkkosivuillaan käytössä olevat aineistot, muuttujat ja luokittelut käyttäen virastojen kehittämiä kuvausjärjestelmiä Meta-Plus (Statistiska centralbyrån, 2008) ja Times (Danmarks Statistik, 2008).

Tilastovirastojen aineistot soveltuvat ensisijaisesti yhteiskuntatieteelliseen tutkimukseen. Terveystieteen tutkijoiden tarvitsemat aineistot eivät ole olleet saatavissa tilastovirastojen etäkäyttäjärjestelmien välityksellä muualla kuin Tanskassa. Tanskassa tilastovirasto on sopinut mm. syntymä-, hoito-, lääke- ja syöpärekisterien kopioiden ylläpidosta, jotta tutkijat voisivat käyttää näissä rekistereissä olevia tietoja yhdessä tilastoviraston tietojen kanssa. Ruotsissa rekisterien kopioimista tilastovirastoon vain tutkimustarkoitusta varten ei pidetä hyvänä toimintatapana.

### 3.3.3. Tietosuoja

Useimmat etäkäyttäjärjestelmien välityksellä yksikötasoista aineistoa tarjoavat maat ovat lähteneet siitä, että paljon muuttujia sisältävää yksikötasoista aineistoa on mahdotonta anonymisoida. Australian valmisaineistot eroavat muiden maiden aineistoista siinä, että Australian tilastovirasto pyrkii anonymisoimaan valmisaineistot ennen niiden tarjoamista tutkijoille. Anonymisointi ei kuitenkaan etäkäyttäjärjestelmässä ole yhtä tiukkaa kuin jos aineisto luovutettaisiin CD:llä.

Sen sijaan, että aineiston tietosuoja pyrittäisiin varmistamaan ennen tutkijoille luovuttamista, pyrkimyksenä on enemmän suojata yksilöiden tiedot muilta ulkopuolisilta. Tavoitteeseen päästään, kun tutkimusprojektin ja luovutuksen eri osa-alueiden varmistetaan olevan mahdollisimman tietoturvallisia. Mitä tarkempia tietoja tutkija saa (eli mitä suurempi todennäköisyys on, että havaintoyksiköitä voidaan tunnistaa aineistosta), sitä tietoturvallisempi pitää aineiston jakelukanavan olla. Englannissa tilastovirasto ja tietoarkisto käyttävät aineiston tarkkuustason ja jakelukanavan arvioinnissa viittä tekijää:

1. projektin luotettavuus

2. tutkijoiden luotettavuus
3. aineiston tietosuojaja
4. jakelukanavan turvallisuus
5. tulosten luovutuksen turvallisuus

Jos yhden kriteerin turvallisuustaso on korkeampi, muiden kohtien turvallisuusvaatimuksia voidaan madal-  
taa. Niinpä täysin anonymisoitu aineisto (turvallinen aineisto) voidaan jakaa verkkosivuilla (ei-turvallinen  
jakelukanava). Vastaavasti, mikäli jakelukanava on suhteellisen turvallinen (etäkäyttö) ja henkilöt ovat tur-  
vallisia (sitoumuksen antaneet tutkijat), aineiston tietosuojaa ei tarvitse turvata niin tarkasti.

Kaikissa tarkastelluissa järjestelmissä tilastoviraston tulisi myös, ainakin satunnaisesti, tarkistaa tutkijoiden  
tulostukset. Ruotsissa ja Tanskassa tilastovirastot tarkistavat aina määrällisesti suuret tulostukset ja tekevät  
muuten vain satunnaisia tarkistuksia. Hollannissa on jouduttu resursoimaan runsaasti työvoimaa tulostus-  
ten tarkastamiseen, ja tästä syystä Hollannissa suositetaan, että tutkijat myös kirjoittaisivat tutkimuksensa  
valmiiksi etätyöpöydällä. Näin yksi tarkastus yhtä projektia kohden olisi riittävä. Australiassa on sen sijaan  
kehitetty automaattisia tarkastusmenetelmiä varsinkin taulukoiden tarkistusta varten.

Käyttäjien tunnistautumisessa ja kirjautumisessa etäkäyttöjärjestelmään on eri maissa eri käytäntöjä. Hol-  
lanissa käyttäjät tunnistetaan sormenjälkitunnisteella. Tanskassa ja Ruotsissa käytetään RSA-avainta yhdes-  
sä pin-koodin kanssa. Englannissa on puolestaan käytössä käyttäjätunnistusefedeeraatio.

### 3.3.4. Järjestelmä

Järjestelmät muistuttavat teknisesti toisiaan pääpiirteiltään. Tanskan järjestelmää on käytetty esikuvana  
muille järjestelmille. Etätyöpöytäratkaisut muistuttavat hyvin paljon omalla työpöydällä työskentelyä. Ka-  
nadan järjestelmä eroaa selkeästi muista tarkastelussa olevista järjestelmistä. Kanadassa on käytössä etä-  
ajoon perustuva järjestelmä, jossa tutkija ei näe yksilötasoista aineistoa, vaan saa tietoa aineistossa olevista  
muuttujista ja muodostaa näiden tietojen perusteella ohjelmakoodin, joka lähetetään etäkäyttöjärjestel-  
mään ajettavaksi. Tulos lähetetään tutkijalle sähköpostitse.

Etäkäyttöjärjestelmissä on tarjolla tavallisimmat tilasto-ohjelmat, kuten SAS, SPSS ja STATA. Mm. Ruotsin  
järjestelmästä on tietoturvasyistä päätetty jättää pois ohjelmointikieliä; tutkijat voivat sen sijaan lähettää  
koodinsa sähköpostitse tarkistusta varten ennen ajoa. Ruotsin tilastoviraston on vaihtanut maksullisista  
Windows Office-ohjelmista maksuttomaan Open Officeen lisenssikustannusten vuoksi. Myös Tanskan tilas-  
tovirasto on siirtynyt kustannussyistä Citrix-järjestelmästä käyttämään Windows Terminal Services (nykyi-  
nen Windows Remote Desktop Services) -järjestelmää. Koska tarjottavien tilasto-ohjelmien lisenssimaksut  
kasvattavat etäkäyttöjärjestelmän ylläpitokustannuksia, Hollannin tilastovirasto on päätenyt tarjoamaan  
SAS-ohjelman käyttöä ainoastaan niille, jotka maksavat siitä erikseen.

### 3.3.5. Kustannukset ja rahoitus

Ruotsin tilastoviraston arvion mukaan etäkäyttäjärjestelmän ylläpito maksaa noin 700 000 euroa vuodessa. Tästä summasta kuluu neljä henkilötyövuotta (noin 400 000 euroa) käyttäjätukeen ja it-järjestelmien ylläpitoon sekä laitteiden, ohjelmien ja lisenssien rahoitukseen (noin 200 000 euroa). Etäkäyttäjärjestelmän kustannuksiin lasketaan myös vanhojen aineistojen kuvailu. Tilastovirasto saa korvausta vuosittain näistä kustannuksista Data Infrastructure committee:lta (DISC). Tutkijat maksavat järjestelmän käytöstä ainoastaan silloin, kun heille tehdään uusia aineistoja. Etäkäyttäjärjestelmän maksuttomuus koskee vain tutkijoita; viranomaiset maksavat etäkäyttäjärjestelmän käytöstä. Ruotsissa etäkäyttökustannuksiin ei lasketa aineistojen suunnittelua ja poimintaa tietovarastosta, koska tämän työn tekee aineistoasiantuntija tilasto-osastolla. Tilastovirasto maksaa aineistoasiantuntijan työn.

Tanskassa aineistojen suunnittelu ja poiminta lasketaan osaksi tilastoviraston ylläpitämää tutkijapalvelua, jonka vastuulla on etäkäyttäjärjestelmän ylläpito. Tutkijapalvelussa työskentelee 14 henkilöä. Vuosikustannukset ovat noin 1,6 miljoonaa euroa, josta puolet katetaan tutkimusministeriöstä saadulla rahoituksella. Tutkijat eivät maksa etäkäyttäjärjestelmän käytöstä mutta kustantavat heille muokattujen tutkimusaineistojen muodostamista..

Hollannissa tutkijat maksavat puolet etäkäyttäjärjestelmästä ja sen ylläpidosta koituvista kustannuksista. Jokaisella aineistolla on tietty hinta, ja tutkimusprojektin liittämistä etäkäyttäjärjestelmän piiriin lasketaan.

Englannin tilastoviraston etäkäyttäjärjestelmä on todettu kustannustehokkaaksi verrattuna tilastovirastojen tutkimuslaboratorioihin, sillä noin 260 000 eurolla pystytään kattamaan 120 projektin ja näissä projekteissa työskentelevien 150 tutkijan palvelut.

## 3.4. Yhteenveto eri maiden ratkaisuksista

Taulukko 3 Eri maiden etäkäyttäjärjestelmien ratkaisut

Toiminta	Ruotsi	Tanska	Hollanti	Englanti	Australia	Kanada
käyttöönottovuosi	2005	2001	2006	2004	2003/2011	2011
etätyöpöytäympäristö	x	x	x	x	x	
eräajoympäristö					x	x
tutkimuslaboratorio			x	(x)	x	x
Citrix järjestelmä		(x)	x	x	-	
Windows TS-järjestelmä	x	x			-	
tutkijoille suunniteltuja valmisaineistoja	(x)	(x)	x		x	x
anonymisoidut aineistot					x	

tulostusten tarkastus	(x)	(x)	x	x	x	x
myös muiden kuin tilastoviraston aineistoja	(x)	x	(x)	x		-
aineistot yhdisteltävissä toisiin	x	x	x			
myös muu kuin tutkimuskäyttö	x			x	x	
ulkopuolinen rahallinen tuki	x	x		x		-
tutkijoilta kerätty rahoitus		1/2	1/2	-	(x)	
käyttäjämäärät 2009	500	650	200	150	800	?

\* - = ei tietoa, x = täysin näin, (x) = enimmäkseen näin

## 4. Hankkeen aikana opittuja asioita

### 4.1. Tutkijoiden näkökulmia

MIDRAS-hankkeen aikana kerättiin eri alojen tutkijoilta näkemyksiä ja mielipiteitä etäkäyttöjärjestelmään liittyvistä mahdollisuuksista, toiveista ja uhista. Tutkijat antoivat myös palautetta rekisteritutkimuksen tämänhetkisestä tilanteesta. Tämä palaute on otettu huomioon MIDRAS-toimintamallia, -järjestelmää ja niiden suosituksia suunniteltaessa.

Hankkeen aikana järjestettiin kaksi muutostarpeita ja kehitysehdotuksia kartoittavaa tutkijapaneelia, joihin osallistui eri alojen ja tutkimuslaitosten rekisteritutkijoita. Projektiryhmä tapasi tutkijoita myös erillisissä tutkijatapaamisissa. Lisäksi hanketta on esitelty rekisteritutkimusaiheisissa seminaareissa.

Tässä luvussa olevan tutkijapalautteen lisäksi luvussa viisi esitellään MIDRAS-pilottijärjestelmää käyttäneiden tutkijoiden näkemyksiä ja palautetta itse pilottijärjestelmästä. Pilottijärjestelmää käyttäneiltä tutkijoilta saatu palaute laajempiin kysymyksiin on yhdistetty tässä luvussa muun tutkijapalautteen yhteyteen. Tutkijoiden ja pilottikäyttäjien palautteissa korostuvat pääosin samat asiat.

#### 4.1.1. Lähtökohta

Tutkijat ovat korostaneet tarvetta uusiin toimintatapoihin rekisteritutkimuksen nykytilanteen ongelmien ratkaisemiseksi. Tutkijoiden näkökulmasta rekisteritutkimuksen ongelmat ovat ennen kaikkea rekisteriaineistojen pitkät toimitusajat, aineistojen anonymisointivelvoite, aineistojen korkea hintataso, hankala lupahakemusprosessi ja aineistojen kuvailujen puutteellisuus. Varsinkin yhteiskuntatieteellisen alojen tutkijoiden mukaan tilastolain alaisia rekisteripohjaisia tutkimusaineistoja pitää saada paremmin käyttöön. Tämän vuoksi tutkijat ovat korostaneet tarvetta tilastolain muuttamiseen.

Tutkijoiden näkökulmasta MIDRAS-etäkäyttöjärjestelmä voisi helpottaa ja tehostaa rekisteritutkimusta. Tämä edellyttää kuitenkin kokonaisvaltaista käytäntöjen muutosta. Etäkäyttöjärjestelmän pitää tutkijoiden mukaan tarjota ratkaisuja myös muihin rekisteritutkimuksen ongelmiin kuin teknisiin kysymyksiin.

Jotta etäkäyttöjärjestelmä toimisi hyvin, on otettava huomioon tutkimusprosessin eri vaiheet. Tutkijat ovat selvityshankkeen aikana toistuvasti korostaneet, että oheispalvelujen (aineistokatalogi, keskitetty lupahakemus) merkitys etäkäyttöjärjestelmässä on vähintään yhtä tärkeä kuin itse etätyöpöytä. Tutkijoiden toive on erityisesti, että etäkäyttöjärjestelmä suoraviivaistaisi, nopeuttaisi ja helpottaisi aineistojen hakemisprosessia.

#### **4.1.2. Tutkijoiden tarpeet ja toiveet**

Tutkijat toivovat, että käytettävissä olisi valmiita, laadukkaita tutkimusaineistoja, joihin saa nopeasti käyttöluvan. Keskitetty aineistokatalogi tai hakupalvelu taas voisi tukea merkittävästi tutkimuksen suunnittelua, tarjoamalla helposti ja kattavasti selkeän kuvan siitä mitä tietoa on saatavilla. Tutkijat ovat ehdottaneet myös esimerkkiaineistojen tuottamista. Näin voitaisiin saada uusia tutkijoita rekisteritutkimuksen piiriin. Kokeneemmat rekisteritutkijat ovat toivoneet pääsevänsä käyttämään primäärirekestereiden raakadataa nykyistä enemmän.

MIDRAS-järjestelmään on myös toivottu entistä laajempia aineistoja, kuten tilastoviranomaisten laajoja kyselyaineistoja. Etäkäyttöjärjestelmän kautta saataviin tutkimusaineistoihin on voitava myös yhdistää tutkijoiden omat aineistot. Muutoin järjestelmä ei palvele tutkimuksen tekemistä.

Tutkijoiden mukaan järjestelmän toimintavarmuus ja käytön vaivattomuus ovat erityisen tärkeitä. Järjestelmän on toimittava eri tilanteissa luotettavasti ja kapasiteetin on oltava riittävä eri aineistojen käsittelemiseen.

Tutkijat ovat toistuvasti painottaneet laadukkaiden metatietojen tärkeyttä. Tutkimusaineistojen hyvät metatiedot ja aineistojen laadukkuus ovat tarpeen, jotta tutkimusprosessi voi edetä ja onnistua. Aineistojen tarkistaminen ja korjailu vie aikaa itse tutkimukselta. Tutkijat ovat myös ehdottaneet, että tutkimusaineistojen käyttökokemukset ja aineiston perusteella tehdyt tutkimukset lisättäisiin osaksi metatietoja. Esimerkiksi toisiaan tukevien muuttujien yhteisestä käyttötarpeesta ei saa tietoa ilman erillistä ohjeistusta.

Varsinaista tutkimusta edeltää pitkä aineiston käsittelyn prosessi, jossa tutkimusjoukon muodostamisen jälkeen tietyillä kriteereillä poimitaan tutkimusaineisto. Tutkimusaineistojen yhdistely, yhtenäistäminen ja esikäsittely voivat olla työläitä prosesseja. Etäkäyttöjärjestelmän tuleekin tukea koko aineistojen käsittelyyn liittyvää prosessia. Esikäsittelyn työläyden vuoksi tutkijat ovat ehdottaneet, että MIDRAS-järjestelmässä tarjottaisiin uudelleen käytettäväksi eri tutkimusprojektien muodostamia ja muokkaamia tutkimusaineistoja. Vähintään toivotaan, että aineistojen esikäsittelyssä käytetyt työkalut ja ohjelmistot tallennettaisiin MIDRAS-järjestelmään, eli osaksi toimintamallia tulisi koodinjakovelvoite.

Tutkijat ovat tuoneet esiin, että tutkimusaineistojen pseudonymisoinnissa henkilötunnuksesta on irrotettava tutkijalle tietoa ennen pseudonymisointia. Tällaisia tietoja ovat esimerkiksi ikä, sukupuoli ja tieto virheellisestä henkilötunnuksesta... Joka tapauksessa tutkijat näkevät tarpeelliseksi tehdä yhteiset pseudonymisointiohjeet.

#### 4.1.3. Haasteita ja uhkakuvia

Tutkijoiden näkökulmasta etäkäyttöjärjestelmään liittyy monia avoimia kysymyksiä ja huolenaiheita, jotka liittyvät niin yhteistyöhön eri toimijoiden välillä, tutkimusaineistoihin kuin etätyöpöydällä työskentelemiseen.

Tutkijat ovat olleet huolissaan siitä, sitoutuvatko eri viranomaiset ja muut toimijat riittävästi järjestelmän kehittämiseen ja ylläpitämiseen. Erityisenä huolena on, jääkö etäkäyttöjärjestelmän ulkopuolelle sellaisia tietoja ja aineistoja, jotka olisivat tutkimuksen kannalta keskeisiä. Eri toimijoiden yhteistyö on tutkijoiden näkökulmasta tärkeää etäkäyttöpalveluiden toteuttamiseksi.

Tutkijat ovat lisäksi korostaneet, että tutkimusaineistojen muodostaminen on vaativa ja pitkä prosessi, jolla on omat erityispiirteensä. Viranomaisilla on paljon kokemuksen tuomaa hiljaista tietoa, jonka avulla he muodostavat tutkimusaineistoja. Jos tutkija poimii itse aineistonsa, hän joutuu tekemään sen ilman vastaavaa tietämystä. Tutkijan vastuun ja toisaalta vallan lisääminen tutkimusaineistojen tuotantoprosessista nähtiinkin hieman kaksitahoisena asiana. Keskeinen kysymys tutkijoiden mielestä on myös se, antaako järjestelmä vastauksia tutkimusaineistojen anonymisointiin liittyviin kysymyksiin, vai jääkö tutkimusaineistojen anonymisointikäytäntö ennalleen.

Tutkijat ovat olleet huolissaan etäkäyttöjärjestelmän vaikutuksesta tutkimusten kustannuksiin. Etäkäyttöjärjestelmän ei tulisi lisätä tutkimusaineistojen hintaa eikä muitakaan kustannuksia, kuten tarvetta hankkia päällekkäisiä lisenssejä tilasto-ohjelmiin. Mahdollisten käyttömaksujen tulisi olla suhteutettu muihin työskentelymahdollisuuksiin. Pilottijärjestelmää kokeilleet tutkijat näkivät ehdottomasti maksuttomina kaiken sen tiedon mitä tutkija tarvitsee suunnitteluvaiheessa: aineistojen selailu, niihin tutustuminen, rekisterien tietosisältöihin ja luokituksiin tutustuminen. Aineistojen kustannukset ovat sidoksissa viranomaisten työhön aineistojen muodostamisessa, joten tämän prosessin toivotaan yksinkertaistuvan ja vaikuttavan näin myös kustannuksiin.

Kustannusten lisäksi tutkijapaneeleissa esitettiin huoli siitä, lisääkö järjestelmä tutkijan työtä ja viranomaisten työkuormaa ja toimitusaikeita. Tällaisessa tilanteessa järjestelmä ei toisi merkittävää parannusta nykytilanteeseen. Lisäksi tutkimusaineiston käsittelytyö ei saisi vaikeutua etätyöpöydän, käyttöliittymän, suorituskäytön tai yhteysongelmien vuoksi. Esimerkiksi tutkimusaineiston koon tai käytössä olevan tietokoneen ei

pitäisi vaikuttaa tutkimusaineiston käsittelyyn etätyöpöydällä. Jos etäkäyttöjärjestelmän myötä tutkimusten tulokset aletaan tarkastaa, tarkastusmenettelyn pelätään muodostuvan uudeksi pullonkaulaksi tutkimusprosessissa.

#### **4.1.5. MIDRAS-toimintamallin mahdollisuudet**

Hankkeen aikana saadun tutkijapalautteen ensisijainen viesti on, että tutkijat ovat innostuneita etäkäyttöjärjestelmän mahdollisuuksista ja sen tuomista parannuksista. Tietosuojaan ja tietoturvaan liittyvien parannusten lisäksi järjestelmä tarjoaa laajemman ja helpomman pääsyn käsiksi aineistoihin. Samalla tutkimuksen tekemistä edistetään ja aineistojen laatu paranee.

Tutkijoiden mukaan yksi järjestelmän suurimpia mahdollisuuksia on se, että kommunikaatio nopeutuu eri osapuolten välillä. Aineistojen säilytys ja uudelleen tarjoaminen tietyn ajan kuluttua tutkimuksen päättymisestä ovat tutkijoiden mielestä varteenotettavia keinoja rekisteritutkimuksen tehostamiseksi. Tutkijoiden näkökulmasta jo rekisteritutkimuksen prosessin nopeutuminen toisi merkittävää hyötyä tutkimuksen kannalta.

## **4.2. Viranomaisten näkökulmia**

Projektin aikana on kartoitettu rekistereitä ylläpitävien viranomaisten näkökulmaa ja mielipiteitä MIDRAS-toimintamallista ja etäkäyttöjärjestelmästä erillisissä tapaamisissa ja projektin aikana sekä pilotin yhteydessä saadusta palautteesta. Viranomaisilta saatu palaute on ollut projektille tärkeää ja se on otettu huomioon järjestelmän suunnittelussa. Viranomaisilta ei ole pyydetty virallisia lausuntoja MIDRAS:sta, joten tässä esitetyt viranomaisten näkemykset tulee nähdä yksittäisten edustajien kommentteina.

### **4.2.1. Lähtökohta**

Rekisteriaineistoja ylläpitävät viranomaiset ovat suhtautuneet lähtökohtaisesti myönteisesti ajatukseen, että rekisteritutkimusta halutaan helpottaa ja tehostaa niin tutkijoiden kuin viranomaisten kannalta. Etäkäyttöjärjestelmään liittyy kuitenkin viranomaisten kannalta vielä monia kysymyksiä, sovittavia asioita ja kehittämiskohtia, ennen kuin tutkimusaineistoja voidaan tarjota tutkijoille etäkäyttöjärjestelmän välityksellä. Viranomaiset kuitenkin näkevät, että etäkäyttöjärjestelmä voisi olla tehokas ja hyödyllinen, ja että tutkimusaineistojen välitys MIDRAS-järjestelmän kautta voisi helpottaa viranomaisten työtä aineistojen muodostamisessa sekä parantaa palvelun laatua. Projektissa kerättyjen kokemusten perusteella eri toimijoiden mielestä Suomeen halutaan saada yksi yhteinen rekisteritietojen etäkäyttöjärjestelmä.

#### 4.2.2. MIDRAS-toimintamallin toteuttamisen edellytykset ja kehittämiskohteet

Viranomaiset ovat suhtautuneet varovaisen myönteisesti ajatukseen, että aineistoja välitettäisiin ja käytettäisiin MIDRAS-järjestelmän kautta. MIDRAS-toimintamalli edellyttää kuitenkin vielä monia päätöksiä ja muutoksia. Viranomaiset ovat kuitenkin ymmärtäneet, että teknisten uudistusten lisäksi myös käytäntöjä pitää muuttaa, jotta rekisteritietojen saatavuus paranee. Viranomaisten tulee analysoida ja muuttaa prosessejaan ja toimintatapojaan. Tämä koskee niin aineistojen tuottamista, toimenkuvia ja vastuita kuin tutkijoiden kanssa toimimistakin.

Viranomaisilta saatujen kommenttien mukaan asenteiden ja totuttujen tapojen muuttaminen on haaste etäkäyttöjärjestelmän käyttöönotolle. Osaamisen kehittäminen vie aikaa. Esimerkiksi lupaprosesseissa ja niiden yhtenäistämässä on paljon kehitettävää. Resursseja vaaditaan ennen kaikkea järjestelmän rakentamis- ja käyttöönottovaiheessa. Prosessien ja tietojärjestelmien kehittäminen sekä metatietotyö ovat rekistereitä ylläpitävillä viranomaisilla vaativia tehtäviä. Tutkimustarkoitustakin tukevan dokumentaation kehittäminen on merkittävä vaatimus ja tutkimusta varten muodostettujen valmisaineistojen tekeminen vaatii työpanosta. Kaikkineen aineistojen kokoaminen tutkimuskäyttöön on suuri haaste, jonka vuoksi lisäresurssointi on tarpeellista.

Valmistelu- ja suunnitteluajan ja kohdennettujen resurssien lisäksi viranomaiset toivovat linjauksia ja päätöksiä uuteen toimintamalliin siirtymisestä sekä yhteisiä toimintatapoja ja metatietojen standardointia. Tiedonsiirroista ja tietoturvaratkaisuista tulee tehdä selkeät sopimukset. Tietosuojakysymykset pitää ottaa yleisellä, yksityiskohtaisella ja lainsäädännön tasolla huomioon.

Rekistereitä ylläpitävät viranomaiset ovat nähneet MIDRAS-järjestelmän viranomaiskäytön hyvänä mahdollisuutena. Järjestelmän kehittäminen ja laajentaminen viranomaisten käytettäväksi on saadun palautteen perusteella kannatettava ja hyvä asia, joka vaikuttaisi myönteisesti myös viranomaisten omaan toimintaan.

#### 4.2.3. Tutkimusaineistojen muodostaminen ja luovuttaminen

Viranomaiset ovat korostaneet, että heidän asiantuntijoillaan on sellaista osaamista tutkimusaineistojen muodostamisessa, johon tutkijoiden tiedot tai taidot eivät välttämättä riitä. Tämän takia onkin tärkeää ylläpitää viranomaisten ja tutkijoiden välistä yhteyttä, jotta olemassa oleva tietotaito saadaan hyödynnettyä. Viranomaiset ovat painottaneet myös kyselyaineistojen merkitystä rekisteriaineistojen rinnalla.

Viranomaistenkin mielestä metatiedot ovat keskeinen osa etäkäyttöjärjestelmän kehitystyötä. Metatietotyöhön vaaditaan ajan ja työpanoksen lisäksi tarkempaa metatietojen määrittelyä ja standardointia. Järjestelmän suunnittelu ja käyttöönotto edellyttävät myös tiedostomuotojen määrittelyä.

Projektissa ei selvitetty yksityiskohtaisesti rekisterien teknisiä talletusmuotoja ja tietokanta-arkkitehtuuria. Viranomaisilta on kuitenkin saatu kommentteja myös aineistojen muodostamisen ja sähköisen toimittamisen toimintatavoista. On selvää, että viranomaisten valmiustaso liittyä rajapintojen kautta MIDRAS-järjestelmään on vielä vähäinen. Siksi aineistojen rajapintapalveluiden käyttöönotossa on edettävä vaiheittain. Tutkijalle luovutettavien aineistojen siirtäminen järjestelmään etukäteen (ns. push-malli) nähtiin toteuttamiskelpoiseksi vaihtoehdoksi jo nykytilanteessa, sen sijaan tietojen kysely rajapintojen kautta (ns. pull-malli) viranomaisten tietokannoista vaatii viranomaisilta kehittämistyötä.

### 4.3. Lakitekniset kysymykset

Projektin järjestämään lakipaneelin osallistui kolmen rekistereitä ylläpitävän viranomaisen lakimiehet sekä tietosuojavaltuutettu. Lakipaneelissa keskusteltiin, millä tavalla uusi järjestelmä vaikuttaa rekisteriaineistojen käyttöön lainsäädännön näkökulmasta.

Koska lait on laadittu teknologiariippumattomiksi, ei lakiasiantuntijoiden mukaan laissa nähdä eroa sille, luovutetaanko aineisto CD:llä vai pääseekö tutkija näkemään yksikkötason aineiston etäkäyttöjärjestelmän välityksellä. Näin ollen eri lakien aineiston luovutusta käsitteleviä kohtia tulee noudattaa siitä huolimatta, että tutkijat eivät saa etäkäyttöjärjestelmän välityksellä fyysisesti kopiota tutkimusaineistosta. Tietosuojavaltuutettu myös muistutti, että yksilötasoinen aineisto määritellään henkilörekisteriksi siinäkin tapauksessa, ettei aineisto sisällä suoria tunnisteita, mutta henkilö on mahdollista tunnistaa aineiston tiedoista tai yhdistämällä henkilötunnisteet takaisin aineistoon koodiavaimella. Tästä syystä etäkäyttöjärjestelmän välityksellä käytössä oleviin pseudonymisoituihin rekisteriaineistoihin tulee suhtautua kuten henkilötunnuksellisiin henkilörekistereihin. Tietojen luovuttamisessa tulee siis noudattaa henkilötietolain mukaista tarkoituksenmukaisuusperustetta: tutkijalle tulee ainoastaan luovuttaa niitä tietoja, joita tutkija on osoittanut tarvitsevansa tutkimuksessaan. Tutkija määritellään hänelle muokatun henkilöaineiston rekisteripitäjäksi, vaikka tutkija ei saisi aineistoa omaan haltuunsa.

Tilastolain alaisten aineistojen luovuttamiseen etäkäyttöjärjestelmän ei nähty tuovan isoja muutoksia nykyisen lainsäädännön ollessa voimassa: aineistot tulee anonymisoida riippumatta luovutustavasta. Tilastoviranomaisen pitää yhdistää tilastolain alaiset aineistot muiden viranomaisten aineistoihin. Näin ollen ainoastaan kuolinsyytiedot sekä tilastolaissa poikkeuksiksi luetellut tiedot ammatista ja koulutuksesta voidaan toimittaa etäkäyttöjärjestelmän välityksellä käytettäväksi samalla tavalla yksilöaineistoina kuin muiden viranomaisten rekisteritiedot.

Lakimiespaneelissa keskusteltiin valtiovarainministeriön asettaman tilastolain uudistamista pohtivan työryhmän esittämistä mahdollisista muutoksista. Epäselvää on, tulisiko EU:n tilastoasetuksessa käytetty termi käyttöoikeudesta rinnastaa etäkäyttöön, tutkimuslaboratoriotyöskentelyyn vai fyysiseen aineiston luovu-

tukseen. Tilastokeskuksen lakimies katsoi, että vaikka tilastolakia muutettaisiin niin, että tutkijat voivat saada käyttöoikeuden anonymisoimattomiin aineistoihin, tämä muutos koskisi ainoastaan uudistuksen jälkeen kerättyjä aineistoja. Tilastokeskuksen mukaan myös tulevaisuudessa aineistojen yhdistäminen tulisi suorittaa Tilastokeskuksessa, mikäli tutkija käyttää useiden viranomaisten tietoja.

## **5. Selvityshankkeen aikana toteutettu pilotti ja siitä saatu palaute**

### **5.1. Pilotin suunnittelu ja tarkoitus**

Selvityshankkeessa toteutettiin pilottijärjestelmä. Pilotin suunnittelun lähtökohtana käytettiin kansainvälisiä rekisteritietojen etäkäyttöjärjestelmiä ja Tilastokeskuksen TK Online -järjestelmää. MIDRAS-pilotissa etäkäyttö ei kuitenkaan ole osa kansallisen tilastoviranomaisen tutkimuspalveluita, vaan erillinen järjestelmä, johon tuodaan tietoja suoraan useammilta viranomaisilta. Pilotin suunnittelussa otettiin huomioon myös tutkijapaneelissa esiin tulleet rekisteritutkijoiden kokemukset ja etäkäyttöjärjestelmään kohdistetut toiveet.

Pilotoinnin tavoitteena oli testata käytännössä ehdotettavia toimintamalleja sekä tietotekniikan että hallinnon osalta. Tarkoituksena oli tuottaa pilottijärjestelmä, joka olennaisilta osin vastaa ehdotettua varsinaista MIDRAS-järjestelmää, ja jonka hallinnollisia ja teknisiä ratkaisuja voitaisiin sellaisenaan soveltaa jatkossa MIDRAS-toimintamallissa. Mikäli pilottivaiheessa valittu ratkaisu osoittautuisi ongelmalliseksi, tavoitteena oli kokemusten perusteella kyetä ehdottamaan vaihtoehtoista ratkaisua.

### **5.2. Pilotin rajaukset ja toteutus**

Jo varhaisessa vaiheessa todettiin, että pilotissa ei voida kokeilla kaikkia suunniteltuja MIDRAS-järjestelmän osia. Varsinainen rekisteritietojen tutkimusetäkäyttö katsottiin pääasiaksi, joten pilotissa toteutettiin käsityönä sellaiset taustapalvelut, joiden on tarkoitus toimia MIDRAS-järjestelmässä automaattisin, vakioiduin tavoin. Näitä palveluita ovat määrämuotoisista metatiedoista koottu aineistokatalogi, siihen perustuva keskitetty lupahakemus, käyttäjähallinnon käyttöliittymät, saatavuus- ja kokeiluaineistot ja tutkimusaineistojen ja -menetelmien säilytyspalvelu.

Pilotin ensimmäisessä vaiheessa testattiin järjestelmän tietointegraatiota ja perustoimintaa pseudodatalla; toisessa vaiheessa pilottijärjestelmää hyödynnettiin varsinaisessa rekisteritutkimuksessa, jossa käytettiin aitoja rekisteriaineistoja.

Pilotin ensimmäinen vaihe toteutettiin CSC:n sisäisesti: tietointegraatio huhtikuussa 2010 ja etätyöpöydän peruskäyttö kesällä 2010. Samaan aikaan etsittiin pilotin tutkimusprojektit ja sovittiin käytännöistä niiden viranomaisten kanssa, jotka toimittivat aineistoja näihin projekteihin. Pilotissa mukana olleita viranomaisia olivat Kansaneläkelaitos Kela, Terveyden ja hyvinvoinnin laitos THL ja Tilastokeskus.

Koska rekisteriaineistojen lupaprosessi kestää nykytilanteessa pitkään ja joka tapauksessa ennustamattoman ajan, pilottia varten etsittiin tutkimusprojekteja, jotka olivat jo saaneet käyttöluvan aineistoihinsa ja joille pilottijärjestelmä voisi toimia aineistojen luovutuskanavana. Toinen vaihe, tutkimuskäyttö, toteutettiin syksystä 2010 lähtien sitä mukaa, kuin tutkimusprojektien aineistot saatiin järjestelmään ja tutkijoille toimitettiin sirukortit ja kortinlukijat, joita tarvittiin etätyöpöydälle kirjautumiseen. Varsinaiset tutkimusaineistot saatiin pilottijärjestelmään syyskuusta 2010 lähtien ja tutkijat pääsivät tekemään tutkimusta pilottijärjestelmän etätyöpöydällä marraskuussa 2010.

Varsinaisen tutkimustyön ja sen teknisten edellytysten lisäksi pilotointiin liittyi paljon kommunikaatiotyötä sekä aineistojen tuottajien että käyttäjien suuntaan. Viranomaiset tutustutettiin MIDRAS-hankkeeseen ja hankkeen yhteydessä toteutettavaan pilottiin, niiden kanssa sovittiin hallinnolliset toimintatavat ja varsinaisen aineistotoimituksen tekniset yksityiskohdat. Lisäksi tutustuttiin tärkeimpien viranomaisten metatietokäytäntöihin.

### 5.3. Pilottijärjestelmä ja sen käyttö

Pilottijärjestelmä oli käytössä 1.7.2010-14.1.2011. Pilotissa toteutettiin suunnitellusta MIDRAS-järjestelmän palveluista seuraavat:

- aineistojen pseudonymisointi
- lähetys ja vastaanotto (osin suunnitellun mallin mukaisesti)
- vahva autentikointi (hyödyntäen mikrosirutunnistautumista)
- etätyöpöytä (Windows remote desktop services -ohjelmistolla)
- käyttäjien työalueet ja analyysi- ja tilasto-ohjelmistot osittain
- sekä aineistojen tarjoaminen tietokannasta.

Pilottijärjestelmää käytti kaksi tutkimusprojektia ja tutkimuksiin käytettiin yhteensä tietoja yhteensä kahdeksasta rekisteristä, kolmelta eri viranomaiselta sekä yhtä tutkijan omaa aineistoa.

Useamman viranomaisen aineistoja käyttäneessä pilottiprojektissa tutkijat yhdistelivät aineistot pseudonymien perusteella etätyöpöydän työkaluilla. Tutkijat käyttivät aineiston käsittelyssä työkaluina SPSS-, R- ja taulukkolaskentaohjelmia sekä aineistotietokantana käytettyä MySQL-tietokantaohjelmistoa.

Tietoturva ja tietosuoja ovat olleet keskeisessä asemassa MIDRAS-ympäristön suunnittelussa. Tietojen päätyminen ulkopuolisten käsiin estettiin hyödyntämällä salakirjoitettuja tietoliikenneyhteyksiä, edellyttämällä käyttäjiltä tunnistautumista vahvalla autentikoinnilla ja eristämällä MIDRAS-palvelu asianmukaisesti.

Viranomaiset pseudonymisoivat aineistot algoritmisesti SHA1-tiivisteillä käyttäen projektikohtaista salalauseetta, joka annettiin aineistojen toimittajille pilottijärjestelmän aineistonvastaanottopalvelun kautta. Ai-

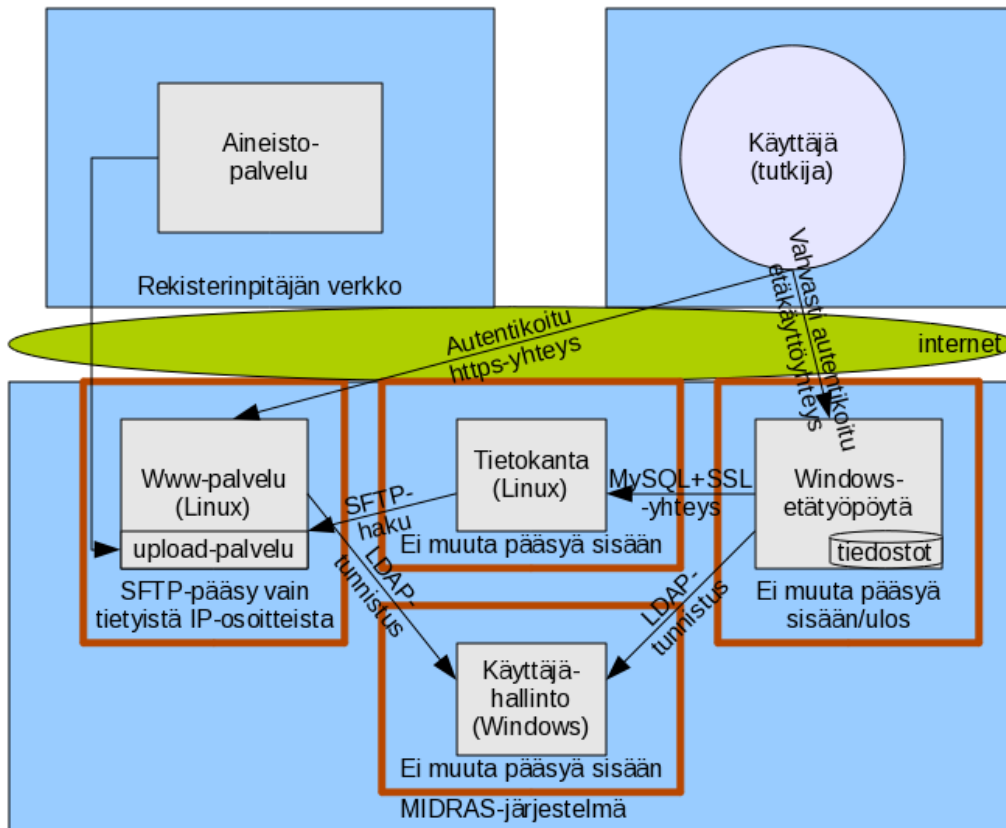
neistot toimitettiin vastaanottopalveluun CSV-muodossa, samoin aineistojen rakenteelliset metatiedot (sarakkeiden nimet ja tyypit). Scp-tiedonsiirtojen tunnistamiseen käytettiin SSH2 protokollan DSA-avaimia. Aineistot annettiin tietokannasta tutkijoiden käyttöön siinä muodossa, jossa viranomaiset olivat lähettäneet ne; aineistoja ei pseudonymisoitu toiseen kertaan. Käyttäjien vahvaan tunnistamiseen käytettiin Väestötietokeskuksen HST-organisaatiosirukortteja. Sirukorteissa käyttäjä tunnistetaan varmenteella, joka saadaan käyttöön sirukortilta PIN-koodilla.

Pilottijärjestelmään käyttäjien tutkijoiden kanssa järjestettiin tapaaminen, jossa näytettiin kirjautuminen etätyöpöydälle sekä ohjeistettiin miten aineistot saa käyttöönsä etätyöpöydällä. Aineistojen kuvailut ja muu tutkijoiden käyttäjätuki hoidettiin sähköpostitse.

#### **5.4. Tietotekninen ympäristö**

Pilottiympäristö toteutettiin neljällä virtuaalipalvelimella, joista kahdessa käyttöjärjestelmänä oli Windows Server 2008r2 ja kahdessa Red Hat Enterprise Linux Server 5.5. Palvelimet olivat omassa aliverkossaan CSC:n testiverkossa. Windows-palvelimet olivat varsinainen etätyöpöytäpalvelin (midrasterm.csc.fi) ja Active Directory -hallintopalvelin (midrasauth.csc.fi); Linux-palvelimet olivat aineistojen säilytykseen käytetty tietokantapalvelin (midrasdata.csc.fi) ja www-palveluita sekä aineistojen vastaanottopalveluita tarjoava apupalvelin (midraswww.csc.fi). Pilotin palvelimet on esitetty kuvassa 1.

Kuva 1 Pilottijärjestelmän palvelimet



Palvelimien tarjoamat palvelut ja niitä varten tehdyt ohjelmistoasennukset on lueteltu taulukossa 4.

Taulukko 4 Pilotin palvelimet, palvelut ja asennukset

palvelin	palvelut	asennukset
midraswww	www-palvelu (http, https), aineistojen vastaanotto (scp)	Apache 2.2.3, PHP 5.1.6, OpenSSH 4.3p2
midrasdata	tietokantapalvelu (MySQL + SSL portissa 3307)	MySQL 5.0.77, stunnel 4.15, aineistojen vastaanottoautomaatti
midrasauth	AD domain controller, autentikointipalvelu, LDAP, SPSS-lisenssipalvelu	SPSS-lisenssipalvelin, Fujitsu mpollux DigiSign
midrasterm	etätyöpöytäpalvelu, analysiohjelmit	Windows Remote Desktop Services, stunnel 4.15, MySQL connector/ODBC, MySQL Workbench, R 2.11.1, RODBC, Python 2.7, PyODBC, SPSS for Windows 19, OpenOffice .org 3.2.1

## 5.5. Pilottijärjestelmän käyttäjäpalaute

Pilottijärjestelmän käyttäjiltä kerättiin palautetta käyttäjätutkimuksen avulla tammikuussa 2010. Tutkimus suoritettiin henkilökohtaisten haastattelujen (2 käyttäjää/Helsinki) ja sähköpostitse tapahtuneen kyselyn (3

käyttäjää/Tampere) avulla. Kaikki tutkimukseen osallistuneet vastasivat samaan avoimeen kysymysrunkoon.

Tavoitteena oli kerätä testikäyttäjiltä palautetta ja kehitysideoita erityisesti MIDRAS-pilottijärjestelmästä sekä myös yleisempää palautetta rekisteriaineistojen etäkäytöstä ja rekisteriaineistojen käyttöön liittyvistä tarpeista. Pilottijärjestelmän käyttäjäkokemus ei ollut siinä mielessä täysin autenttinen, että käyttäjät olivat hakeneet luvat aineistoihin normaalin nykykäytännön mukaisesti ja aineistojen vastaanotto järjestelmään sekä tulosten siirto pois järjestelmästä toteutettiin hieman eritavoin kuin varsinaisessa etäkäyttöjärjestelmässä on suunniteltu. Käyttäjille tarjottiin myös erittäin paljon henkilökohtaista neuvontaa, apua ja opastusta; kirjallisia ohjeita ei juuri ollut. Pilottikäyttäjiltä saatiin kuitenkin arvokasta palautetta ja ideoita järjestelmän jatkokehitystä varten. Palautetta käydään läpi seuraavissa kappaleissa, käyttäjien suorat kommentit on merkitty kursivilla.

### 5.5.1. Yhteenveto

*”Olisi kauhean kätevää jos yhdestä palvelusta pääsisi moniin aineistoihin... Jos etäkäyttösystemiin saisi yhdistettyä useita rekisterejä se olisi mahtavaa! Jos kaikki tiedot olisi saatavilla samasta paikasta ja jos jotenkin pääsisi käsiksi ilman kamalan pitkää luparumbaa. Siinä olisi todella hienot mahdollisuudet tutkijalle!”*

Kokemukset järjestelmän käytöstä olivat pääosin positiivisia. Käyttäjät kokivat etäkäyttöpalvelun toivottavaksi, tutkijoita hyödyttäväksi palveluksi, joka voisi edistää ja monipuolistaa rekisteritutkimuksen tekoa ja parantaa sen laatua.

*”Nopeasti tottuu, kätevää, että saa aineistoon yhteyden olinpaikasta riippumatta. Tilasto-ohjelmien löytyminen työpöydältä on hyvä juttu. Lisäksi aineiston muovattavuus on nopeaa ja joustavaa. Usean eri tahon tilastot saadaan kätevästi samaan paikkaan samalle työpöydälle”*

Etäkäyttöpalvelun koettuja hyötyjä rekisteriaineistojen käyttäjälle, verrattuna nykytilanteeseen olivat erityisesti seuraavat:

- Avoimuuden ja tietoisuuden lisääminen. Metatietojen näkyvyyden paraneminen, rekisteritietojen olemassaolon ja tietosisältöjen parempi tunnettuus
- Vaivattomuus päästä käsiksi laajempiin aineistoihin. Rekisterien käytön kynnyksen madaltuminen ja käytön monipuolistuminen
- Aineistojen saatavuuden helpottuminen (ei enää muistitikkuja ja cd-romeja)
- Aineistojen yhdistely eri lähteistä vaivattomampaa
- Parempi tietoturva
- Aineistojen päivityksen/korjaamisen nopeutuminen

- Tilastovälineet helposti saatavilla.

MIDRAS-pilottijärjestelmää pidettiin pääosin helppokäyttöisenä ja tehokkaana. Vaikeuksia aiheuttivat kuitenkin yhteyden epävarmuus ja erityisesti sirukortin käyttöön liittyvät asennustarpeet. Vahvan tunnistautumisen menetelmäksi valittu sirukorttitunnistautuminen osoittautuikin hankalaksi ratkaisuksi.

*”Helppo se oli, helpompi kuin olin kuvitellut”*

*”Alkuun vaikeuksia sirukortin toiminnassa, saatuani uuden kortin ja lukijan, homma toimi”*

Käyttäjät näkivät selkeitä hyötyjä siinä, että aineistojen käsittely tapahtuu etätyöpöydällä eikä omalla koneella:

- parempi tietoturva
- parempi datan säilytysvarmuus, varmuuskopiot
- aineistojen päivitys ja täydentäminen voi onnistua nopeasti ja vaivattomasti
- enemmän tehoa ja käsittelymahdollisuuksia (esim. prosessori useammalla ytimellä, päällekkäisten simulaatioiden mahdollisuus) kuin omalla koneella käsitellessä.

Haittapuolena nähtiin se, että käsittely vaatii aina toimivan nettiyhteyden.

*”Jos koneet hajoaa, on varmuuskopiot siellä. Ja ei tarte stressata henkilötiedoista ym. kun ne tiedot ei ole omalla koneella, on ikävä olo jos ne on omalla koneella ja sitä vaikka kuljettelee.”*

*”Aina ei pääse etätyöpöytäjärjestelmään, niin samat aineistot pitäisi olla myös läppärillä.”*

Jatkokehityksen kannalta mahdollisuuksia lisäarvon tuottamiseen nähtiin erityisesti järjestelmän vuorovai-  
kutteisuuden vahvistamisen kautta.

*”Yhteydenpito helpottui toisen projektilaisen kanssa, yhteisen projektitilan kautta. Tähän vuorovai-  
kutuksen kehittämiseen kannattaisi panostaa enemmänkin.”*

Käyttäjien mukaan ideaalinen etäkäyttöpalvelu olisi helppokäyttöinen, joustava, kattava (koskien niin tarvittavia aineistoja kuin ohjelmia, sisältäen runsaasti tietoa rekistereistä ja luokituksista), vuorovai-  
kutteinen, nopea ja tehokas.

*”Hinta, laatu ja käytettävyys on kuljettava käsi kädessä”*

Jatkokehityksessä tulee huomioida myös tavoiteltujen käyttäjäryhmien erilainen tietotekninen osaamistaso ja kokemus analysointiohjelmien käytöstä. Järjestelmän tulee tukea eri käyttäjäryhmien tarpeita, osaamis-

tasoa ja työskentelyä. Varsinkin satunnaisille käyttäjille selkeä ohjeistus eri osa-alueiden käytön tueksi on hyvin tärkeää.

Etäkäyttöpalvelun kehittämisessä erityisen tärkeitä huomioitavina korostettuja seikkoja:
<ul style="list-style-type: none"> <li>• Toimintavarmuus</li> <li>• Käyttömahdollisuus eri koneilla, ”milloin ja missä haluaa”</li> <li>• Kattavuus, niin aineistojen kuin ohjelmistojen ja työkalujen suhteen</li> <li>• Aineistojen kattavat kuvailutiedot helposti saatavilla</li> <li>• Tehokkuus ja nopeus</li> <li>• Helppokäyttöisyys ja tarvittava ohjeistus helposti saatavilla</li> </ul>

### 5.5.2. Käyttäjäpalautetta pilottijärjestelmän eri osiin liittyen

Seuraavaan taulukkoon on koottu tarkempaa käyttäjäpalautetta ja kehitysideoita järjestelmän eri toimintoihin liittyen.

Taulukko 5 Yksityiskohtaisempaa käyttäjäpalautetta

Palaute	Käyttäjäkommentit
<b>Yhteyden muodostaminen (Remote desktop connection tai rdesktop –ohjelmalla)</b>	
<p>Yhteyden luominen etätyöpöytään koettiin pääosin helpoksi.</p> <ul style="list-style-type: none"> <li>- Käyttäjät eivät joutuneet asentamaan itse etäkäyttöohjelmaa ja sen käyttöönottoon annettiin opastusta.</li> </ul>	<p><i>”Helppoo se nyt oli, mut jos en olis aiemmin käyttänyt olis varmaan joutunut vähän miettimään.”</i></p>
<p>Yhteyden muodostaminen oli kuitenkin osalle käyttäjiä ollut hidasta tai epävarmaa.</p> <ul style="list-style-type: none"> <li>- Tämä aiheutti epä tietoisuutta.</li> </ul>	<p><i>”Hidas, joka aiheutti epätietoisuutta aluksi, onko yhteys löytymässä vai ei...”</i></p>
<p>Suurin ongelma oli yhteyden muodostumisen epävarmuus tai pätkiminen. Yksi käyttäjästä ei pystynyt lainkaan käyttämään järjestelmää omalla koneellaan.</p> <ul style="list-style-type: none"> <li>- Useita syitä (sirukortit, palomuurit) joista ei täyttä varmuutta</li> </ul>	<p><i>”Suurin ongelma oli juuri se, että kun omalta koneelta pääsy ei onnistunut. Aika on tutkimuksessa olennaista ja rajallista, on tärkeää että sinne pääsee silloin kun itse haluaa. Nyt se ei vaan onnistunut, ei se sinänsä hankalaa olisi, periaatteessa yksinkertaista jos vaan olisi toiminut.”</i></p> <p><i>”Jos oikeasti olisi pitänyt tehdä töitä niin, että yhteys jatkuvasti pätkii ja joutuu odottamaan katkeako se vai ei, niin ei kenelläkään ole sellaiseen aikaa oikeasti pittemällä tähtäyksellä.”</i></p>
<p><b>Kehitysehdotuksia:</b></p> <ul style="list-style-type: none"> <li>- Jatkoa ajatellen myös etäyhteysohjelman käyttöönotto vaatii selkeän kirjallisen ohjeistuksen.</li> <li>- Etäyhteyden muodostamiseen ja käyttöön liittyvät epävarmuudet tulee selvittää. Mahdolliset epävarmuustekijät, niiden syyt ja keinot ehkäistä tulee selvittää myös käyttäjille.</li> </ul>	

<b>Kirjautuminen ja tunnistautuminen (mikrosirukortin avulla)</b>	
<p>Mikrosirutunnistautuminen osoittautui käyttäjien kannalta hankalaksi ratkaisuksi. Suurin huono puoli mikrosirutunnistautumisen käytössä on vaadittavat erillishankinnat ja asennukset.</p> <ul style="list-style-type: none"> <li>- Prosessi on työläämpi ja hitaampi kuin muiden tunnistautumistapojen kanssa.</li> <li>- Myös oikeanlaisten sirukorttien saamisessa oli hankaluuksia ja se oli vienyt aikaa.</li> </ul>	<p><i>”Toimivan sirukortin saanti ja kortinlukijan hankinta viivästyttivät aloitusta varaamalla tutkimusvapaalla”</i></p> <p><i>”Mikrosiru voi olla ihan kätevä, jos sillä on sitten muutenkin käyttöä, esim. työaikaseuranta ja jos se on jo muutenkin käytössä. Mutta kun käytin vaan pari viikkoa, niin joku salasanalista olisi voinut olla kätevämpi, joku muu tapa voisi olla nopeampi ottaa käyttöön kun se ei vaadi mitään erillistä tilaamista tai asentamista.”</i></p> <p><i>”Pankkitunnukset toimisi missä vaan ja ilman eri ohjelmia. Universaali käyttö on olennaista. Sirukortti sitoo käytön vain tiettyihin koneisiin ja palomuuriongelma voi olla useassa organisaatiossa. Ja se vaatii ohjelmia erikseen.”</i></p>
<p>Käyttäjät olisivat myös kaivanneet etukäteen lisätietoja siitä mitä asennuksia mikrosirukortin käyttö edellyttää.</p>	<p><i>”En ole aiemmin käyttänyt henkilökortin lukijaa, hetken jouduin etsimään ohjelmaa asennettavaksi. Se asennuslinkki kyllä sitten löytyi ihan hyvin. Mutta sitä en tajunnut että joutui erikseen asentamaan ajurit USB lukijalle, odotin että se olisi toiminut automaattisesti. Se että ajurit joutui erikseen asentamaan oli hankalaa, tai ei hankalaa, mutta en tiennyt että se täytyy tehdä niin ihmettelin miksei toimi. Asennukseen tarttisi tietoa, että mitä tarvitsee asentaa”</i></p> <p><i>”Pitäisi etukäteen tietää mitä tarvitaan yhteyden saamiseksi”</i></p>
<p>Käyttäjillä ei ollut alkuasennusta jälkeen ongelmia mikrosirukortin käytön kanssa, joskin sen suhde yhteysongelmiin ja yhteyden muodostumisen hitauteen oli hieman epäselvää.</p>	<p><i>”Tarvittavia kortinlukijoita on vaikeaa löytää kaupoista. Muuten ihan ok.”</i></p>
<p><b>Kehitysehdotuksia:</b></p> <ul style="list-style-type: none"> <li>- Tunnistautumismenetelmää valittaessa tulee luotettavuuden lisäksi painottaa myös käyttöönoton vaivattomuutta ja nopeutta, toimintavarmuutta, kustannuksia sekä mahdollisuutta käyttää järjestelmää helposti myös useammalta koneelta</li> <li>- Käyttöönoton tulee olla mahdollisimman nopeaa ja vaivatonta, niin ettei se vaadi useita asennuksia ja hankintoja.</li> </ul>	<p><i>”Helppo saatavuus (on tärkeää).”</i></p>
<b>Omien aineistojen hakeminen järjestelmästä, vastaanotto etätyöpöydälle, tarkastaminen ja kuvailutiedot</b>	
<p>Käyttäjät olivat hakeneet aineistot suoraan käytettävään ohjelmaan/avanneet ne ohjelmalla, ja aineistojen sijainti järjestelmän hakemistossa oli jäänyt epäselväksi.</p>	<p><i>”Datan osoitteen olisin tarvinnut. Analyysit ja data oli nyt eri kansioissa, ei ollut ihan selkeätä mikä on missäkin. Ja vähän koko järjestelmästä ja sen sisällöstä, nyt oletin että projektin sisällä kumpikin pääsisi käsiksi samoihin kansioihin, muttei oikein selvinnyt oliko niin.”</i></p> <p><i>”Etätyöpöydällä saisi olla jonkin tyyppinen ilmoitustaulu</i></p>

	<i>tai ohjekansio, mistä data löytyy.”</i>
Aineistojen kuvailutiedoissa oli ollut puutteita. Toisaalta osalle käyttäjiä etäkäyttöjärjestelmä oli mahdollistanut nopean tietojen täydentämisen.	<i>”Jokin informaatio datan sisällöstä olisi ollut tarpeen.” ”Aineiston metatiedot oli aika hankala saada ja koota, eikä niitä aina saatukaan. Keräilin niitä pyytämällä aineiston lähettäneiltä yhteisöiltä.” ”Puutteita (kuvailutiedoissa) oli aluksi, etätyöpöytä mahdollisti nopean korjaamisen.”</i>
<b>Kehitysehdotuksia:</b> <ul style="list-style-type: none"> <li>- Aineistojen mukana on tarpeellista saada mahdollisimman kattavat kuvailu- ja taustatiedot. Ainakin datojen ja muuttujien sisällöt, selitykset, mahdolliset arvot ja vaihteluvälit, luokitukset ja niiden selitykset, selitys kuinka koodaus on tehty. Kuvailutietojen kattavuus on hyvin tärkeää niin aineistojen tarkastamiselle kuin käsittelylle.</li> <li>- Kuvailutietojen esittämisessä olennaista on, että kuvailutiedot on löydettävissä ja tarkastettavissa helposti ja nopeasti samalla kun käsittelee itse aineistoja.</li> <li>- Esitystapaan ei ollut selkeitä ehdotuksia. Pari käyttäjistä piti hyvänä erillistä koodikirjaa järjestelmässä.</li> </ul>	<i>”Nyt aineistot oli tutut, mutta yleensä rekisteriaineistot sisältää aina paljon vieraita muuttujia ja datan tarkastaminen on hyvin tärkeää... erityisesti jos tiedot on yhdistetty monesta rekisteristä.” ”Olennaista on, että samalla pitäisi pystyä tarkastelemaan sekä dataa että kuvailutietoja, yhtäaikainen tarkastelumahdollisuus on tärkeää.” ”(Yleensä käyttämissäni aineistoissa) seuraa mukana erillinen tiedosto, se ei kyllä ehkä ole kauhean näppärää. Esim. SPSS:ssä taustatiedot saa selville hiirtä tiedon kohdalla klikaten, saa muuttujan taustatiedot, ei tarvitse avata erillistä tiedostoa tai skrollata ja etsiä oikeita tietoja, että missä kohtaa ne on... Toisaalta jos aineisto on SPSS:ssä niin kuvailutietoja on sieltä vaikea saada muuhun ohjelmaan, lisäinformaatio usein häviää. Pitäisi olla sitten myös erillinen tiedosto, jossa on samat asiat.”</i>
<b>Aineistojen käsittely etätyöpöydällä</b>	
Aineistojen käsittely etätyöpöydän kautta ei käyttäjien mielestä juuri eronnut omalla koneella työskentelystä.	<i>”Ei eroa perusohjelmien käytöstä kun oli tuttu aineisto.” ”Isommalla aineistolla järjestelmän käyttö voisi olla tehokkaampaa. Nyt se oli aika samanlaista peruskäyttöä, kuin olisi ollut omalla koneella.”</i>
Aineistot olivat pääosin käyttäjien tilaamassa ja tarvitsemassa muodossa, tosin aineistojen ja muuttujien tarkempia kuvailutietoja olisi kaivattu. Muutamia pieniä ongelmia oli päivämäärien ja ääkkösten kanssa. Pilottikäyttäjät eivät kokeneet, että järjestelmä olisi tarjonnut erityistä apua aineistojen tulkinnalle.	<i>”Ainoastaan ääkköset vaati kikkailua. Jotain tapahtui ääkkösille. Omalla koneella Ä ja Ö:t näkyi normaalisti, mutta etäkäyttöjärjestelmässä ne oli muuttuneet kirjainjonoiksi” ”Päivämäärätietojen kanssa tuli ongelmia, kun ne olivat päivämäärämuodossa vaikka päivätieto puuttui. SPSS ei osannut tulkita tieto päivämäärätiedoksi vaan hylkäsi kaikki tiedot.”</i>
Etätyöpöydän etuna nähtiin se, että sen kautta tilasto-ohjelmat olivat helposti saatavilla. Toisaalta ongelmia oli aiheuttanut tarvittavien työkalujen puuttuminen ja yhteysongelmat.	<i>”Ongelmat tulivat lähinnä siitä, että pilottiympäristössä ei pystynyt tekemään kaikkea mitä olisi oikeasti töitä tehdessään ollut pakko tehdä... Excel tai jokin taulukkolaskentaohjelma, jolla saman vastaajan eri vastaukset pystyy erottelemaan on tarpeen, koska tilasto-ohjelma ei siihen kykene.” ”Suojattu yhteys pitäisi saada kaikkialta, yhteys ei saisi pätkiä ja jos dataa ei saa osissakaan ulos, niin kaikki tarvittavat ohjelmat yms. pitää olla käytettävissä, jotta etätyöpöytä houkuttelee asiakkaita järjestelmälle.”</i>
Myös ohjeistusta ja opaskirjoja kaivattiin.	<i>”Nyt kun ei päässyt esim. nettiin samalla niin tilasto-</i>

	<i>ohjelmien opaskirjan pitäisi siksi olla ohjelman sisällä, että sitä voisi käyttää samalla.”</i>
Tulosten tallennuksessa ja uudelleen käyttöönnotossa ei ollut suuria ongelmia. Yksi käyttäjä olisi kaivannut selkeämpää ohjeistusta siihen, miten pitää toimia jos työskentelyssä pitää pidemmän tauon ja jotta työskentelyn jatkaminen olisi mahdollisimman toimivaa.	
<b>Kehitysehdotuksia:</b> - Koska ajatuksena on, että aineistojen käsittely tapahtuu täysin etätyöpöydällä, varsinaista palvelun kehitystä ajatellen on tärkeää, että se sisältää kaikki aineistojen käsittelyssä tarvittavat ohjelmat ja työkalut ja/tai niiden lisääminen omalle työpöydälle on helppoa.	<i>”Jotain tilasto-ohjelmia varmasti olisi lisättävä jos etätyöpöytä olisi ainoa työväline; esim otoskokolaskenta, epidemiologiset testit jne.” ”Olisin kyllä tarvinnut odotetun eliniän taulun, ohjelmanpätkän, se olis sinne sitten pitänyt lisätä. Mutta onko se ajatus, että siihen järjestelmään voisi lisäillä omia juttuja? Olisi kyllä hyvä jos voisi lisätä omia koodinpätkiä.” ”En paljon katsellut mitä siellä oli tarjolla. Itse olen käyttänyt myös tekstieditoria, UltraEdit tms. ja sitten jotain muuta tapauskohtaisesti.”</i>
<b>Tulosten siirtäminen itselle</b>	
Käyttäjät olivat yrittäneet siirtää aineistoja ja tuloksia sekä koodinpätkiä monin tavoin omalle koneelleen ja olisivat myös halunneet tulostaa niitä kesken käsittelyn. Käyttäjille oli myös hieman epäselvää, että kaiken työskentelyn olisi tarkoitus tapahtua etätyöpöydällä.	<i>”Tulostamismahdollisuus omalla koneella puuttui ainakin alussa ja sitä kaipasin...taulukoiden tallennuksen omalle koneelle olisi hyvä tapahtua heti.” ”En tiennyt, tai ehkä mulla ei vaan ollut joku päällä, siis etten pystynyt ottamaan esim, koodin pätkää tai muuta pois copy pastena sieltä etätyöpöydältä omalle koneelleni. Se ei toiminut tai ehkä olis pitänyt vaan ruksata joltain... leikepöydän kautta ei voinnu kopasta aineistoja jatkokäsittelyyn omalle koneella tai vaikka printattavaksi.”</i>
Koska järjestelmä ei sisältänyt kaikkia tarvittavia työkaluja, ajatus siitä, että käyttö tapahtuu ainoastaan etätyöpöydällä, eikä aineistoja tai tuloksia voi ottaa välillä jatkokäsittelyyn omalle koneelle vaikutti hieman hankalalta.	<i>”Vaikeudet tulivat lähinnä siinä, että kyseessä on pilotti ja oikeat työt eivät suju; esim. tuloksia ei saanutkaan ulos millään tavalla.”</i>
Varsinaista tulosten siirtämistä hankaloitti yhteyden pätkiminen.	<i>”Datojen siirron yhteydessä yhteyden jatkuva pätkiminen hidasti töitä suunnattomasti. Tulostiedostot ehdottomasti olisi saatava vaivattomasti omalle koneelle.</i>
<b>Kehitysehdotuksia:</b> - Käyttäjiiä tulee ohjeistaa selkeästi, että kaikki aineistojen käsittely tapahtuu etätyöpöydällä, eikä aineistoja tai mitään osaa saa poistettua. - Tämä tulee varmistaa myös teknisesti. - Tulosten siirtämisen on tärkeää toimia varmasti, luotettavasti ja nopeasti.	<i>”Ainakin jos sähköpostilla on lähetelty aineistoja niin ne ei sitten ole aina auennut oikeassa muodossa. Pitää olla testattu, varma systeemi, ettei tulokset huku tai aukea eri muodossa.” ”Resurssien hallinnan kautta jos toimisi, siihen olen totunut. Ja pienet välitulokset voisi siirtää raakamuodossa ihan copypastena.”</i>
<b>Yleinen käyttäjäkokemus sekä MIDRAS-järjestelmän toteutus ja toiminta</b>	
Yleisesti ottaen käyttäjät pitivät MIDRAS-järjestelmää	<i>”Varsin yksinkertainen järjestelmä” ”Helpompi kuin ajattelin”</i>

<p>sekä tehokkaana että helppona niin ottaa käyttöön kuin käyttäinkin.</p> <p>Järjestelmän käyttö edellyttää kuitenkin tiettyjä tietoteknisiä perustaitoja.</p>	<p><i>"Kuin olisi omaa konetta käyttänyt kun data oli vaan valmis. Haaste on varmasti siinä, että saa varmistettua datat ja yhdistettyä ne"</i></p> <p><i>"Vaatii kohtalaisen hyviä tietojenkäsittelytaitoja"</i></p>
<p>Vaikeuksia aiheutti pääosin asennus- ja yhteysongelmat.</p>	<p><i>"Ihan kuin omalla koneella olisi tehnyt töitä, lukuun ottamatta viimeisen päivän datojen siirron yhteydessä yhteyden jatkuva pätkiminen hidasti töitä suunnattomasti."</i></p> <p><i>"Kirjautuminen oli hankalin, muuten toimi tosi hyvin"</i></p> <p><i>"(Oli tehokas) muuten, mutta jos yhteys pätki niin meinasimme mennä hermot odotellessa yhteyden löytymistä"</i></p>
<p>Käyttäjät kokivat, että he olivat saaneet reilusti etukäetietoa ja opastusta.</p>	<p><i>"Sain hyvin tietoa. Mutta jos järjestelmä tulee oikeasti käyttöön niin sitten se vaatii käyttöohjekirjan tai oppaan."</i></p>
<p>Erityisesti uudenlainen interaktiivisuus ja yhteistyömahdollisuudet projektin sisällä nähtiin erityisen positiivisena järjestelmän mahdollistamana asiana.</p>	<p><i>"Jaettu hakemisto, ainakin periaatteessa olisi ollut mahdollista tallennella sinne tuloksia ja koodinpätkiä, ja toinen projektilainen olisi sitten voinut tsekkailla, kun yleensä niitä pitää lähettellä sähköpostilla edes taas. Mutta ei oikein opittu käyttämään sitä mahdollisuutta täysin, se jäi vähän kesken."</i></p>
<p><b>Kehitysehdotuksia:</b></p> <ul style="list-style-type: none"> <li>- Varsinainen järjestelmä tarvitsee tuekseen selkeät, helposti saatavilla olevat kirjalliset askelittaiset ohjeet eri käyttövaiheisiin ja toimintoihin. Käyttäjät tarvitsevat selkeän kirjallisen ohjeistuksen myös koko järjestelmän käyttöönottoon ja vaadittaviin asennuksiin liittyen.</li> <li>- Järjestelmän interaktiivisuuden ja sosiaalisten kommunikaation tuen edelleen kehittäminen tarjoaa paljon mahdollisuuksia</li> </ul>	<p><i>"Mitään ohjeita ei järjestelmässä ole. Ne olisivat varmaan tarpeen pitemmän käyttötaujan jälkeen."</i></p> <p><i>"Pieni opas voisi olla alkuun ja ohjeet miten käytetään, ja mistä löytää luokitukset. Perusohjeet perustoimintoihin."</i></p> <p><i>"Ohjeistusta tarvitaan työpöydälle."</i></p>
<p><b>Jatkokehitysideoita</b></p>	
<p>Keskeistä on kattava aineistokokoelma ja tutkijalle entistä parempi mahdollisuus päästä näkemään aineistoja (esimerkiksi demojen avulla).</p>	<p><i>"Tutkijalle tulisi olla nykyistä laajempi pääsy tai näkymä aineistoihin, jotta konkretisoituu mitä tietoja on tarjolla... Pitäisi olla vaikka demoja joita voisi tarkastella etukäteen eri aineistoista."</i></p>
<p>Myös pilottikäyttäjät korostivat sähköisen lupahakemuspalvelun tärkeyttä.</p>	<p><i>"Miten saataisiin valtavaa lippubyrokratiaa vähennettyä tutkijalta?"</i></p> <p><i>"Kaikki palvelut samalla hakemuksella kiitos"</i></p>
<p>Vuorovaikutteisuuden edelleen vahvistaminen</p>	<p><i>"Lisäksi siinä voisi olla joku keskustelualusta muille tutkijoille, muillekin kuin oman projektin tutkijoille. Sellainen neuvonta ja keskustelumahdollisuus. Tämä voisi olla hyvä tutkimuksen ohjaajille, ne voisi tarkistaa järjestelmän kautta miten hommat etenee ja mitä on tehty."</i></p> <p><i>"En tiedä onko toteutettavissa vielä interaktiivisemmin päästä ohjeistamaan ja neuvomaan toista projektin sisällä. Että pääsisi hetkellisesti käyttämään toisen etätö-pöytää tai näkemään sen ja toimimaan siellä, yhtäaikai-</i></p>

## 5.6. Viranomaisten edustajien kokemuksia pilotista

*”(aineistojen toimittaminen verkon kautta voisi) Helpottaa teknisesti tietopyyntöjen toteuttamista. Tulevaisuudessa voisi jopa vähentää aineistojen muodostamisen tarvetta meillä (jos tutkija voisi muodostaa aineistot etätyöpöydällä itse), mikä olisi tervetullut helpotus nykyiseen työtilanteeseen.”*

Myös pilottijärjestelmään aineistoja toimittaneiden viranomaisten edustajat antoivat palautetta ennen kaikkea järjestelmän teknisestä toteutuksesta ja aineistojen sähköisestä toimittamisesta.

Pilotin aikana pääosin organisaatioiden palomuurit ja tietoturva-vaatimukset aiheuttivat haasteita tietojen sähköiselle siirrolle. Tässä oli huomattavia eroja eri organisaatioiden välillä: kahden organisaatioiden kanssa aineistot saatiin siirrettyä pienien alkuvaikeuksien jälkeen, yhden organisaation kanssa taas jouduttiin lopulta toimimaan niin, että aineistot siirrettiin perinteisesti CD-levyllä. Kuitenkin myös tämän organisaation edustaja koki, että tekninen tiedonsiirto olisi toteutettavissa, kunhan laajemmat päätökset järjestelmään osallistumisesta olisi tehty.

Viranomaisten edustajat pitivät MIDRAS-toimintamallia, järjestelmän toteutussuunnitelmaa ja tiedonsiirron prosesseja pääosin toimivina, selkeinä ja toteuttamiskelpoisina, myös teknisesti. Järjestelmän ja aineistojen sähköisen siirron kehittäminen palveluksi vaatisi kuitenkin vielä sekä tarkempaa suunnittelu- ja määrittelytyötä että myös erilaisia sopimuksia. Keskeisiä tarkemmin määriteltäviä ja sovittavia asioita olisivat ainakin sopimukset tiedonsiirrosta ja tietoturvaratkaisuista, palomuurit ja muut suojaukset, tekniseen yhteentoimivuuteen liittyvät seikat sekä siirrettävien tietojen ja metatietojen tarkempi määrittely. Aineistojen siirtäminen ns. push-mallin mukaisesti nähtiin suhteellisen helposti ja suoraviivaisesti toteutettavana, kun taas pull-mallin nähtiin vaativan huomattavasti enemmän valmistelutyötä ja muutoksia toimintamalleihin. Pull-mallin mukaista aineistojen jakelua ei nähty lyhyellä aikavälillä toteuttamiskelpoisena.

*”Tuotantosovelluksessa pitää enemmän kiinnittää huomiota eritoten metatietojen määrittelyyn, standardointiin, siirtoformaattiin ja käytettävyyteen”*

*”Push-toiminnallisuus ei edellytä merkittäviä (teknisiä tai yhteentoimivuuteen) liittyviä lisävaatimuksia... Jos taas ajatellaan pull-tyyppistä aineiston jakelua, niin siihen varmastikin liittyisi paljonkin vaatimuksia. Tällä hetkellä en näe sitä toteuttamiskelpoisena lyhyellä aikavälillä.”*

## 6. Käyttäjä- ja järjestelmävaatimukset ja suunnittelua ohjaavat periaatteet

Seuraavassa taulukossa on tiivistelmä selvityshankkeen aikana opituista ja korostuneista MIDRAS-järjestelmän suunnittelua ohjaavista periaatteista. Perusvaatimukset on luokiteltu kolmitasoisella tärkeysluokituksella:

- 1: järjestelmä ei toteuta tarkoitustaan, mikäli vaatimusta ei täytetä
- 2: järjestelmän hyödyllisyys vaarantuu, mikäli vaatimusta ei täytetä
- 3: vaatimus tekee järjestelmästä hyödyllisemmän

Tutkijoiden keskeinen vaatimus on, että järjestelmä helpottaa rekisteritutkimuksen tekemistä. Viranomais-ten keskeisimmät vaatimukset ovat, ettei järjestelmä hankaloita resurssitilannetta ja etteivät tietoturva ja -suoja heikkene.

Taulukko 6 Kooste etäkäyttöjärjestelmän suunnittelua ohjaavista periaatteista

Vaatimuksia tutkijoiden näkökulmasta		Tärkeys
Toimintavarmuus	Järjestelmän ja yhteyksien tulee toimia varmasti ja luotettavasti. Työtä ei saa hukkua eikä työskentely vaikeutua järjestelmän ongelmien takia.	1
Kattavuus aineistojen ja niiden kuvailujen suhteen	Järjestelmän kautta pitää saada mahdollisimman kattavasti aineistoja. Aineistojen kuvailutietojen tulee olla kattavasti, helposti ja nopeasti saatavilla.	1
Lupahakemusprosessin selkeys	Järjestelmän tulee sisältää myös keskitetty sähköinen lupahakemuspalvelu, joka helpottaa ja tehostaa lupahakemusprosessia.	2
Joustavuus	Järjestelmän tulee olla käytettävissä helposti, useammalla koneella, ajasta ja paikasta riippumatta.	2
Helppokäyttöisyys	Järjestelmä ei saa vaatia useita tai hankalia asennuksia ja hankintoja. Mahdolliset yhteensopivuus- ja palomuuriongelmat tulee selvittää ja tiedottaa etukäteen. Järjestelmän tulee tukea ja opastaa käyttäjää niin käyttöönotossa kuin käytön eri vaiheissa.	2
Kattavuus käytössä olevien ohjelmistojen ja työkalujen suhteen	Koska aineistoja ei saa ulos järjestelmästä ja kaikki käsittely tapahtuu järjestelmän sisällä, järjestelmän tulee sisältää kaikki aineistojen analysoinnissa tarvittavat ohjelmistot ja työkalut, sekä tarjota keino niiden ja tutkijoiden omien aineistojen lisäämiseksi.	2
Tehokkuus ja tuloksellisuus	Järjestelmän avulla työskentelyn ja palvelinten tulee olla tehokkaita ja nopeita.	2
Hinnoittelun kilpailukykyisyys	Järjestelmän käytön hinnoittelun tulee olla perusteltua, eikä se saa johtaa "tuplamaksuihin". Hinnoittelun pitää rohkaista käyttämään järjestelmää.	2
Vuorovaikutteisuus	Järjestelmän tulee tarjota lisäarvoa nykyiseen työskentelyyn verrattuna edistämällä sekä projektin sisäistä että laajemminkin tutkijoiden ja viranomaisten	3

	välistä vuorovaikutusta.	
Vaatimuksia rekistereitä ylläpitävien viranomaisten näkökulmasta		Tärkeys
Tietoturvallisuus	Rekisteritietojen tulee olla vain käyttöluvan saaneiden hallussa.	1
Mukautuvuus	Eri lakien erityisvaatimukset aineistonluovutuksiin on otettava huomioon.	1
Yhteensopivuus	Järjestelmän pitää pystyä vastaanottamaan aineistoa riippumatta viranomaisen tietoteknisistä ratkaisuista ja tasosta.	2
Vuorovaikutteisuus	Järjestelmän tulee edistää tutkijoiden ja viranomaisten vuorovaikutusta.	2
Kustannustehokkuus	Aineistojen toimittaminen järjestelmän välityksellä ei saa olla vaikeampaa tai enemmän kustannuksia vaativaa kuin tällä hetkellä Pitkällä tähtäimellä järjestelmän tulee johtaa kustannussäästöihin.	2
Rahoitus	Järjestelmän toteuttamisesta, ylläpidosta sekä oheispalveluiden tuottamisesta ei saa koitua lisäkustannuksia rekistereitä ylläpitäville viranomaisille.	2

## 7. Suositeltu MIDRAS-toimintamalli ja etäkäyttöjärjestelmä

Tässä luvussa kuvataan MIDRAS-toimintamalli ja etäkäyttöjärjestelmä palveluineen sellaisina, kuin ne parhaiten palvelevat suomalaista rekisteritutkimusta sekä aineistojen hyödyntäjien että rekisterinpitäjien näkökulmasta. MIDRAS-järjestelmä ja sen toimintamalli perustuvat usean toimijan yhteistyöhön, joten suunnittelussa on pyritty ottamaan eri sidosryhmien tarpeet tasapuolisesti huomioon. Näin on voitu rakentaa sellainen toimintamalli, johon kaikki toimijat voivat sitoutua ja josta kaikki toimijat saavuttavat konkreettisia hyötyjä.

Luvussa kuvataan ensiksi MIDRAS-järjestelmän toimintaidea ja sen tarjoamat palvelut. Tämän jälkeen esitellään keskeiset prosessit ja toimijoiden vastuut. Lopuksi luvussa kuvataan MIDRAS-järjestelmän tekninen toteutus.

### 7.1. MIDRAS-järjestelmän yleiskuvaus

MIDRAS-järjestelmä on ensisijaisesti tarkoitettu toimimaan kanavana, jonka kautta luovutetaan tutkimuskäyttöön viranomaisrekistereistä muodostettuja yksikkötason tutkimusaineistoja. Käytännössä järjestelmä kuitenkin soveltuu yleiskäyttöiseksi alustaksi, jolla käyttöluvan saaneet voivat tutkia ja analysoida käyttöluvan mukaisia aineistoja saamatta niitä omaan haltuunsa.

Tutkijan kannalta ehdotetun MIDRAS-järjestelmän keskeisimpiä palveluita ovat (kts. kuva 2):

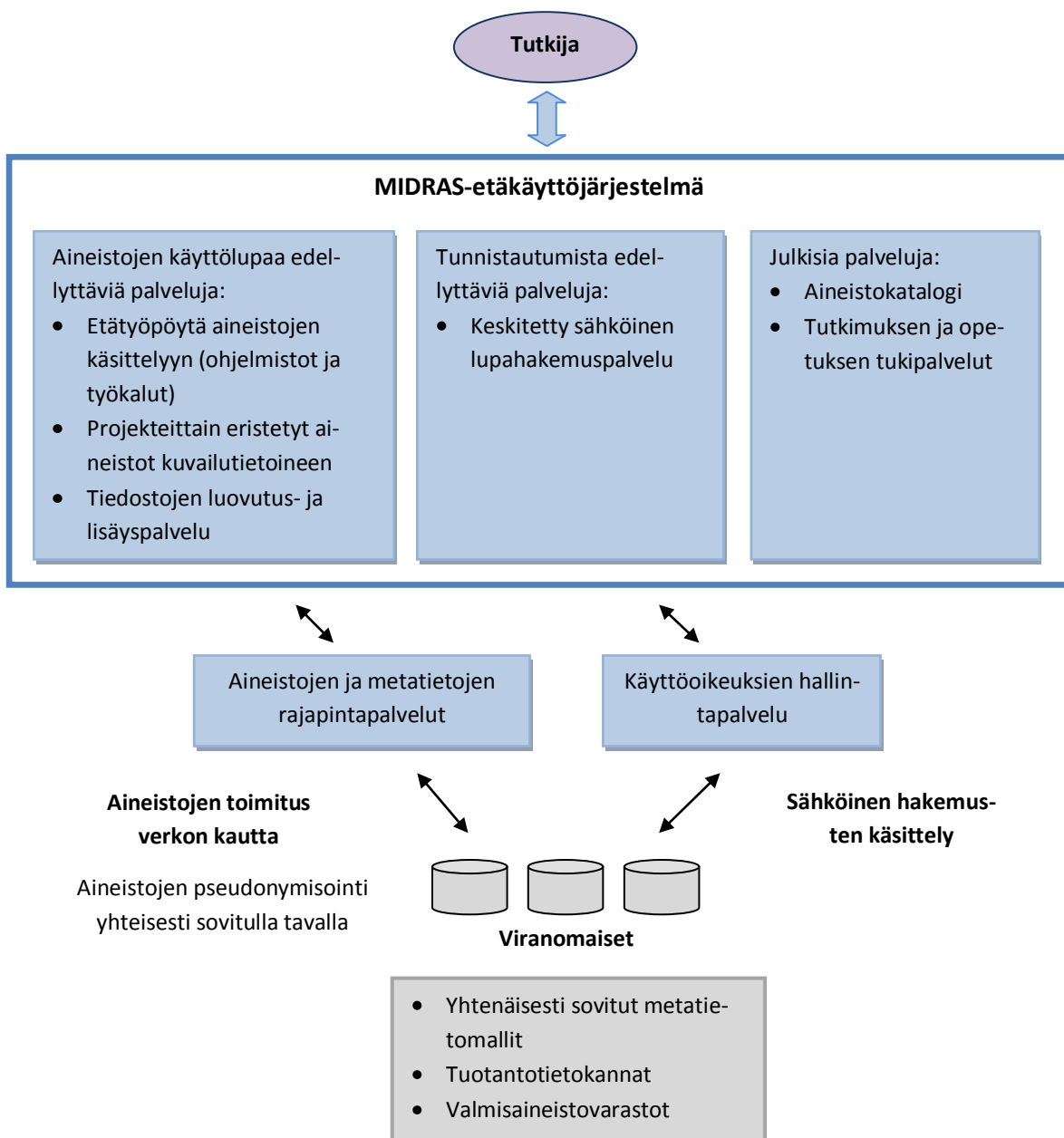
1. Tutkimusaineistojen käsittelyyn ja analysointiin tarkoitettu eristetty etätyöpöytäympäristö palveluineen. Etäkäyttöjärjestelmässä tutkija ei saa tutkimusaineistoa haltuunsa, vaan tutkija saa aineiston käyttöönsä etätyöpöydälle, johon hän voi kirjautua omalta työpisteeltään.

2. Tutkimuslupien hakemista ja tutkimusten suunnittelua helpottava aineistokatalogi ja sille perustuva keskitetty lupahakemuspalvelu. Aineistokatalogi on www-palvelu, johon kerätään mahdollisimman yhteismitallisia metatietoja järjestelmän kautta saatavilla olevista aineistoista.

Aineistoja tuottavan viranomaisen kannalta järjestelmän keskeisimpiä palveluita ovat:

1. Tietojen siirto- ja käyttörajapinta MIDRAS-ylläpito-organisaation ja viranomaisen välillä. Tietojen käyttörajapinnan kautta kulkevat järjestelmän aineistot, metatiedot ja muut tiedot.
2. Käyttöoikeuksien hallinnointityökalu. Käyttöoikeuksien hallinnointityökalu sallii rekisterinpitäjien hallinnoida ja tarkastella, millä tutkimusprojekteilla ja keillä tutkijoilla on pääsy mihinkin aineistoon.

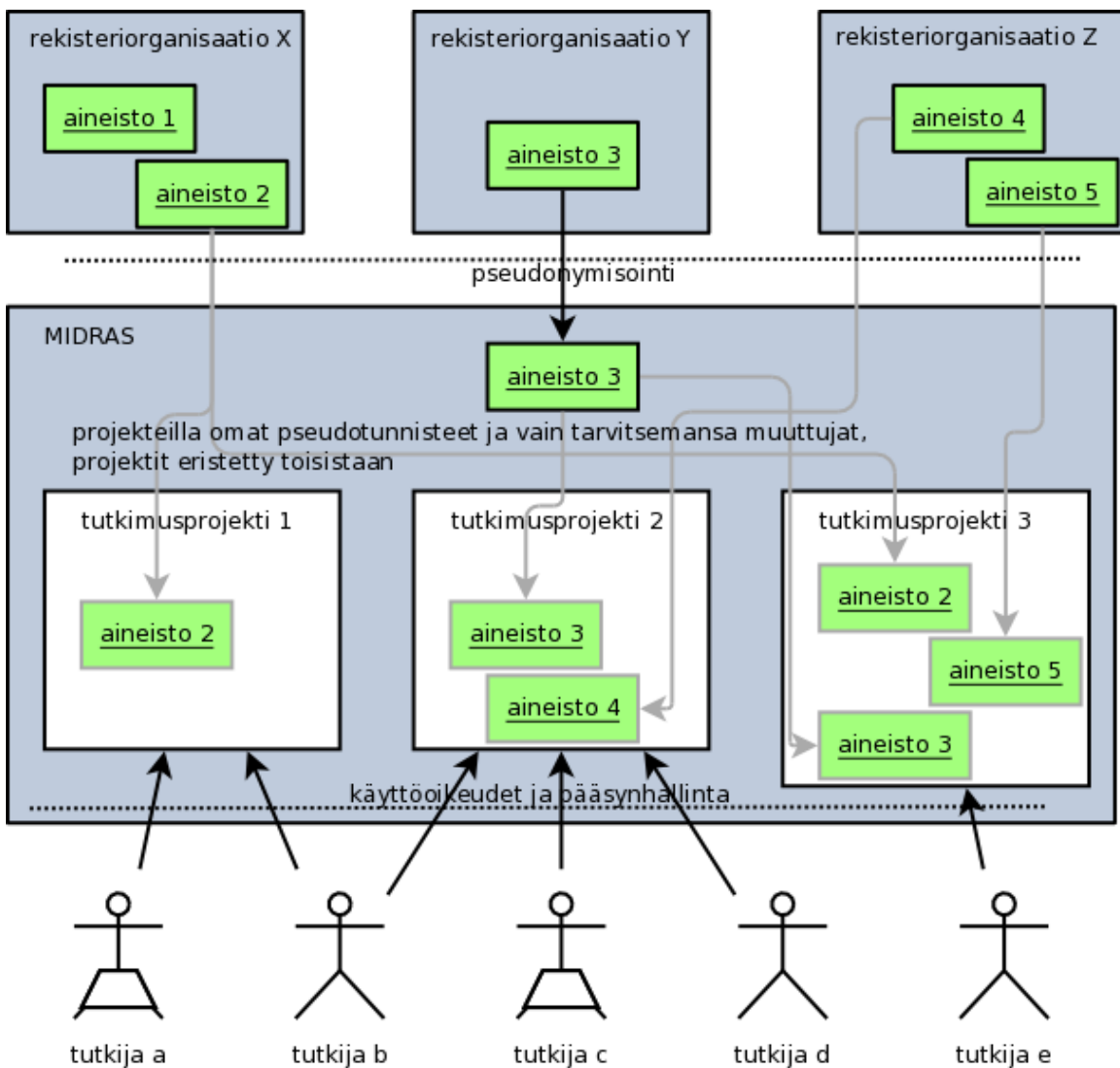
Kuva 2 Kuvaus MIDRAS-toimintamallista ja sen palveluista



MIDRAS-järjestelmässä tutkijat löytävät tutkimustaan hyödyttävät aineistot aineistokatalogin perusteella ja hakevat niihin käyttöluvat keskitetyllä sähköisellä lupahakemuksella. Aineistot toimitetaan MIDRAS-järjestelmään yhteisin käytännöin pseudonymisoituna ja tutkijat analysoivat aineistot etätyöpöydän työkaluilla.

Tietoturva ja -suoja ovat keskeisessä asemassa MIDRAS-ympäristön suunnittelussa. Tietojen päätyminen ulkopuolisten käsiin estetään hyödyntämällä salakirjoitettuja tietoliikenneyhteyksiä, edellyttämällä käyttäjiltä vahvaa tunnistautumista ja eristämällä MIDRAS-palvelu asianmukaisesti (kts. kuva 3).

Kuva 3 Aineistojen suunniteltu kulku MIDRAS-järjestelmässä



## 7.2. MIDRAS-järjestelmän palvelut

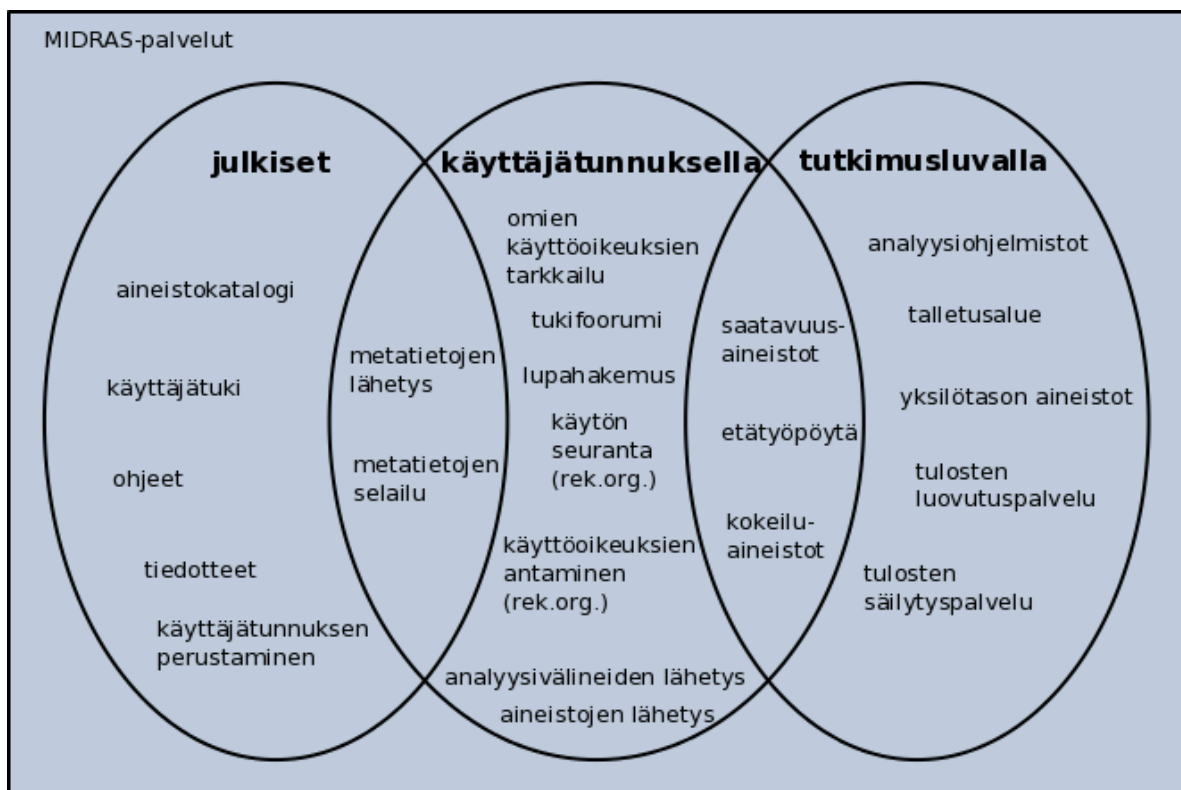
Tässä osiossa kuvataan palvelut, jotka MIDRAS-järjestelmä tarjoaa sekä tutkijoille että aineistojen tuottajille. Samalla määritellään merkittävimmät edistysaskeleet, jotka MIDRAS-toimintamalli tarjoaa suomalaisen

rekisteritutkimuksen tehostamiseksi. MIDRAS-järjestelmän palvelut (kuva 4) voidaan jaotella seuraaviin kokonaisuuksiin:

1. Aineistokatalogi
2. Aineistojen ja metatietojen lähetyksen ja vastaanoton rajapintapalvelut
3. Käyttölupien haku- ja hallintapalvelut
4. Tutkimus-, raportointi- ja analyysiohjelmistopalvelu (etätyöpöytä)
5. Omien tiedostojen lisäys- ja luovutuspalvelu tarkastuksen kautta
6. Tutkimuksen ja opetuksen tukipalvelut

Kuvassa 4 on luokiteltu palvelukokonaisuudet sen mukaan, millaisia käyttöoikeuksia ne vaativat. Esimerkiksi tulosten luovutuspalvelu on luonnollisesti vain niiden käyttäjien saavutettavissa, joilla on voimassa olevat tutkimusluvut aineistoihin. MIDRAS-järjestelmä sisältää kuitenkin myös palveluita, joiden hyödyntämiseen ei tarvita lainkaan käyttäjätunnusta.

Kuva 4 MIDRAS-järjestelmän palvelut



### 7.2.1. Aineistokatalogi

Sisältö	Käyttäjät	Hyödyt	Käyttötapa	Toteutusvaihe
---------	-----------	--------	------------	---------------

MIDRAS-järjestelmän kautta saatavilla olevien aineistojen kuvaukset julkaisu- ja ohjeviitteineen	Tutkijat	Tutkimusten suunnittelu, aineiston käytön tuki, laadukkaat käyttöluvahakemukset	Selailu ja tekstihaut www-selaimella	Perustamisvaihe, palveluvaihe
--	----------	---	--------------------------------------	-------------------------------

Aineistokatalogi on MIDRAS-järjestelmän näkyvin julkinen palvelu. Aineistokatalogin kautta ovat saatavilla MIDRAS-järjestelmään vietyjen aineistojen yksityiskohtaiset kuvailut (metatiedot), joita ei ole erikseen kielletty julkaisemasta. Aineistokatalogin avulla käyttäjät voivat tunnistaa oman tutkimuksensa kannalta merkittävät ja mielenkiintoiset aineistot ja edetä näitä koskevien käyttöluvien hakemiseen. Aineistokatalogi on siis suunnattu erityisesti tutkijoille, mutta aineistokatalogista hyötyvät myös aineistojen tuottajat aineistojen tullessa tunnetummiksi, tutkijoiden käyttöluvahakemuksien selkeytyessä ja aineistojen metatietojen laadun parantuessa käyttäjäpalautteen kautta.

Aineistokatalogin peruskäyttö on aineistojen metatietojen selailua. Jokaiselle sarjalle, aineistolle, muuttujaryhmälle, muuttujalle, koodistolle, koodille ja käsitteelle on oma kuvailunsa, jotka sisältävät myös linkitykset niihin liittyviin kuvauksiin. Aineistokatalogista voi etsiä tietoja rakenteellisella haulilla.<sup>1</sup> Käytössä on myös vapaa tekstihaku, joka etsii hakukohdetta kaikista aineistokatalogin tiedoista. Aineistokatalogia voi käyttää anonyymisti tai sisäänkirjautuneena, jolloin tietoyksiköistä (mm. aineistot ja muuttujat) voi tehdä muistiinpanoja (annotaatioita). Muistiin merkittyihin asioihin voi myöhemmin palata ja niitä voi käyttää pohjana esimerkiksi käyttöluvahakemuksessa.

Aineistokatalogin rakenne perustuu määrämuotoisille aineistokuvailuille, joista katalogin sisältö tuotetaan automaattisesti. Ylläpidon rooli on tarjota seuraavassa kappaleessa kuvattavaa metatietopalvelua, joka auttaa aineiston tuottajia aineistokatalogiin yhteensopivien metatietojen laadinnassa.

Aineistokatalogi on olennainen palvelu rekisteriaineistojen käyttöasteen ja tunnettavuuden lisäämisessä. Aineistojen käyttö tutkimuksissa sisältyy aineistojen metatietoihin, ja aineistojen hyödyntämistä voi seurata sitä kautta. Palvelu on mahdollista toteuttaa vaiheittain siten, että ensimmäisessä vaiheessa kuvataan vain järjestelmään siirretyt aineistot muttei muita tutkimuskäyttöön soveltuvia aineistoja.

### 7.2.2. Aineistojen ja metatietojen rajapintapalvelut

Sisältö	Käyttäjät	Hyödyt	Käyttötapa	Toteutusvaihe
Ohjeet ja menetelmät tietojen pseudonimisoimiseksi ja välittämiseksi MIDRAS-järjestelmään	Viranomaiset	Aineistojen ja metatietojen yhteensopivuus, tietotoimitusten nopeus ja vaivattomuus	Ohjeet ja työkalut haettavissa www-sivuilta, web service -rajapinnat	Perustamisvaihe, jatkokehitysvaihe

<sup>1</sup> esimerkiksi on mahdollista etsiä kaikista aineistoista muuttujat, joiden kuvauksessa on sana "tulot", jotka ovat numeerisia ja joiden kattavuus aineistossa on yli 70%

Rajapintapalveluihin kuuluu varsinainen aineistojen ja metatietojen vastaanottopalvelu MIDRAS-järjestelmässä, sekä tukipalvelut viranomaisille MIDRAS-järjestelmään yhteensopivien aineistojen metatietojen laadinnassa ja rajapintojen käyttämisessä ja toteuttamisessa.

Aineistojen vastaanottopalveluilla viranomaiset voivat avata haluttuja aineistoja tutkimusprojektien käyttöön, päivittää, korjata ja poistaa aineistoja suoraan MIDRAS-järjestelmässä. Palveluun kuuluu myös vastaanotto- ja kyselyrajapintojen dokumentaatio sekä esimerkkityökaluja aineistojen pseudonymisointiin, lähettämiseen, aineistojen muodostamiseen käyttöluvahakemusten perusteella sekä koko prosessin automatisointiin.

Metatietojen vastaanottopalvelulla voidaan siirtää valmiit kuvailut osaksi aineistokatalogia. Palveluun sisältyy myös julkisen, MIDRAS-järjestelmään soveltuvan aineistokuvailun dokumentaation (ml. XML-skeema) ja työkalujen ylläpito. Aineiston kuvailumallia kehitetään aktiivisesti yhteistyössä aineistojen tuottajien ja hyödyntäjien kanssa. Osana palvelua MIDRAS-järjestelmä osallistuu rekisteriaineistoille sopivien metatietokuvauksien kansalliseen ja kansainväliseen kehittämiseen ja standardointiin, koska tämä lisää yhteensopivuutta pitkällä aikavälillä. MIDRAS-järjestelmä osallistuu myös tietojen vaihtoon aineistokatalogien välillä.

Palvelu on aineistojen tuottajien käytettävissä yhteydenotolla MIDRAS-järjestelmän ylläpitoon. MIDRAS-järjestelmän ylläpito myös toimii aktiivisesti metatietopalvelun tehtävissä. Metatietojen kehittäminen on yksi ylläpidon keskeisimmistä tehtävistä. Metatietopalveluun liittyvä työ tulee käynnistää viimeistään järjestelmän käyttöönottoaiheessa.

### 7.2.3. Käyttöluvien haku- ja hallintapalvelut

Sisältö	Käyttäjät	Hyödyt	Käyttötapa	Toteutusvaihe
Aineistojen käyttöoikeuksien haku, myöntö ja tarkastelu	Tutkijat, viranomaiset	Hakuprosessin helppous, nopeus ja läpinäkyvyys; käyttöluvien seuranta	WWW-lomakkeet	Perustamisvaihe

Palvelukokonaisuuteen kuuluvat kaikki käyttöoikeuksiin ja -lupiin liittyvät palvelut. MIDRAS-järjestelmän käyttäjätunnistus kannattaa liittää tarjolla oleviin käyttäjäfederaatioihin kuten suomalaisten korkeakoulujen HAKA-federaatioon. On kuitenkin nähtävissä lukuisia tilanteita, joissa MIDRAS-järjestelmän tunnustettava käyttäjä ei kuulu tällaiseen tunnustusfederaatioon. Tätä varten MIDRAS-järjestelmä sisältää www-lomakkeen, jolla voi perustaa itselleen käyttäjätunnuksen MIDRAS-järjestelmään. Lomakkeella kerätään käyttäjän tiedot ja lisäksi käyttäjän pitää todistaa identiteettinsä esimerkiksi VETUMA-tunnistuksella.

Aineiston käyttöluvahakemus täytetään tunnistetusti sisäänkirjautuneena. Lomakkeeseen syötetään tiedot, joita aineistojen omistajat tarvitsevat käyttöluvan myöntämiseen ja tutkimukseen luovutettavien aineistojen muodostamiseen valmisaineistoista. Tutkijoille pitää olla mahdollisimman helppoa tehdä sellainen lu-

pahakemus, jonka käsittely on vaivatonta myös aineistojen tuottajille. Niinpä lupahakemuksen kehitystyössä on tärkeä panostaa sellaisen käyttöliittymän kehittämiseen, joka tukee laadukkaiden lupahakemusten tekemistä, sekä siihen, että aineistoista on saatavilla sellaisia metatietoja, joiden perusteella tutkijat pystyvät lupahakemuksessaan mahdollisimman tarkasti määrittämään tarvitsemansa aineiston.

Lupahakemuksessa määritellään tutkimuksen yleinen kuvaus, tutkimuksen vastuhenkilö, muut mukana olevat käyttäjät, taustaorganisaatiot ja rahoittajat, tutkimuksessa tarvittavat tiedot, kuvaus siitä, miten tietoja käytetään tutkimuksessa ja varsinainen tutkimussuunnitelma. Jos tutkija yhdistelee tutkimuksessa rekisteriaineistoihin itse keräämäänsä aineistoa, oman aineiston tiedot, kohteiden suostumukset ja keruutapa pitää myös määrittää. Tutkijoiden tiedot saadaan käyttäjätiedoista, jotka saadaan käyttäjätunnistusfederaation kautta tai jotka on syötetty MIDRAS-käyttäjätunnusta perustettaessa. Lisäksi tutkijat tekevät sähköisen tietojen salassapitositoumuksen lupahakemuksen yhteydessä.

Siinä lupahakemuksen osassa, jossa määritellään tutkimuksessa käytettävät tiedot ja niiden käyttötarkoitus, tiedot valikoidaan aineistokatalogista. Jokaisen valitun tiedon ohessa määritellään, mihin tätä tietoa käytetään tutkimuksessa. Lupahakemuspalvelu auttaa tutkijoita valikoimaan tietoja ja esimerkiksi ehdottaa aineistokuvailujen perusteella tutkijan mahdollisesti tarvitsemia lisätietoja.

Käytettävien tietojen määrittelyn osana rajataan tutkittavat havaintoyksiköt (eli määritetään tutkimuskohortti).<sup>2</sup> Yksinkertaisissa tapauksissa tutkija määrittelee muuttujakohtaisesti rajausehdot, joilla tutkimuksen tarvitsemat tiedot poimitaan kokonaisaineistosta. Monimutkaisissa tapauksissa, joissa poiminta tehdään monen eri tiedon välisten suhteiden perusteella, tutkija määrittelee, mitä tietoja tarvitsee kohortin muodostamiseen ("poiminta-aineisto") ja mitä tietoja taas tutkimuskohortin yksilöistä varsinaista tutkimusta varten. Tutkija määrittelee tutkimuskohortin myöhemmin MIDRAS-etätyöpöydän työkaluilla ja saa kohorttimäärittelyn ja alkuperäisen lupahakemuksen perusteella varsinaisen tutkimusaineistonsa (Kohortin muodostamisesta tarkemmin kappaleessa ks. osio "pseudonymisointi ja yhdistely").

MIDRAS-järjestelmän ylläpito huolehtii hakemuspalvelun kehittamisestä ja käyttöluviin perustuvien käyttöoikeuksien kirjanpidosta. Hakemuspalvelua kehitetään helppokäyttöiseksi yhdessä tutkijoiden kanssa.

#### 7.2.4. Tutkimus-, raportointi- ja analyysiohjelmistopalvelu

Sisältö	Käyttäjät	Hyödyt	Käyttötapa	Toteutusvaihe
Aineistojen käsittely ja	Tutkijat	Monipuolinen tutkimusympäristö, yhteistyöalusta, hyvät tietojenkäsittely-	Etätyöpöytä, jolla aineiston käsittelyyn sopivat	Perustamisvaihe

<sup>2</sup> esimerkiksi "kaikki 15-64-vuotiaat henkilöt" tai "lääkeostotapahtumat 2003-2008, joissa ostaja on alle 50-vuotias nainen".

tutkimus		resurssit, tietoturva, tietosuojat	ohjelmistot	
----------	--	------------------------------------	-------------	--

MIDRAS-järjestelmän keskeinen palvelu on etätyöpöytä, jolla tutkija voi käsitellä aineistoja tietoturvallisesti ja tutkimustarpeitaan vastaavasti. Etätyöpöytä tarkoittaa, että tutkija ottaa yhteyden erilliselle palvelimelle, josta käynnistyy tutkijan näytölle työpöytä. Tutkija käyttää etätyöpöydän ohjelmia samalla tavalla kuin oman paikallisenkin koneensa ohjelmia, mutta ohjelmat toimivat oikeasti etäpalvelimella, eivät tutkijan omalla koneella. Tietojen siirto etätyöpöydän ja oman työkoneen välillä on estetty. Tutkimustulokset saa työpöydältä erillisen tulosten luovutuspalvelun kautta.

Etätyöpöytäpalvelimille asennetaan ohjelmat, joita tutkijat tarvitsevat tutkimuksensa tekemisessä. Windows-etätyöpöydällä näitä ovat ainakin tilasto-ohjelmat SAS, SPSS, Stata ja R, ohjelmointikielet Perl, Python ja Java, toimisto-ohjelmat MS Office ja OpenOffice.org, kehitystyökalut Eclipse, IDLE, ja Notepad++ ja aineistotietokannan kanssa yhteensopiva SQL-työkalu. Tutkijat vastaavat tutkimusongelmien selvittämisestä aineistoja analysoimalla. MIDRAS-hallinto vastaa työkalujen tarjoamisesta tutkijoiden käyttöön ja ylläpidosta etätyöpöydällä. Mikäli tutkija tarvitsee työskentelyynsä työkaluja tai tiedostoja, joita ei ole saatavilla etätyöpöydällä, on niitä mahdollista siirtää etätyöpöydälle omien tiedostojen lisäysohjelman kautta.

### 7.2.5. Omien tiedostojen lisäys- ja luovutuspalvelu

Sisältö	Käyttäjät	Hyödyt	Käyttötapa	Toteutusvaihe
Tutkimustulosten luovutus etätyöpöydältä tutkijalle, omien tutkimusaineistojen liittäminen rekisteriaineistoihin, työkalut ja päivitykset	Tutkijat	Tutkimuksen tulosten raportointi, kehitystyön hyödyntäminen ympäristöjen välillä	Hakemisto etätyöpöydällä, www-palvelu	Perustamisvaihe

MIDRAS-järjestelmässä tarvitaan myös keino siirtää tietoja ja tiedostoja etätyöpöydän ja muun maailman välillä. Siirrettäväksi hyväksyttäviä tiedostoja ovat esimerkiksi tutkimusten julkaisuissa käytettävät tulokset, aineistojen käsittelyssä käytettävät työkalut ja anonyymit tilastoaineistot, joita käytetään tutkimuksessa vertailuaineistona. Näiden tietojen siirtäminen tutkimusprojektin etätyöpöydälle on tietosuojan kannalta ongelmaton, kunhan niissä ei ole mitään, mitä voitaisiin yhdistellä tutkimuksessa käytettäviin aineistoihin; siirto etätyöpöydältä pois on ongelmaton, kunhan tiedostoissa ei ole mitään yksilötietoja tai muita arkaluontoisia tietoja.

Tulosten luovutuspalvelu ja tiedostojen lähetys- ja vastaanottopalvelu ovat tiedostovarastoja, jotka ovat etätyöpöydällä saatavilla yhtenä hakemistona käyttäjän kotihakemistossa, ja ulkopuolelta saavutettavissa www-palvelun kautta. Www-palvelusta voi ladata tiedostoja tai julkaista niitä MIDRAS-järjestelmän foorumilla ja uusia tiedostoja voi lähettää palveluun. Kun tiedostot synkronoidaan etätyöpöydän ja www-palvelun välillä, ne läpäisevät tarkastusmenettelyn ja ne talletetaan myöhempää seuranta varten.

### 7.2.6. Tutkimuksen ja opetuksen tukipalvelut

Sisältö	Käyttäjät	Hyödyt	Käyttötapa	Toteutusvaihe
käyttäjätuki, tiedotus- ja keskustelualueet, dokumentti- ja välinevarastot	Tutkijat, viranomaiset	Tietojen, menetelmien ja kokemusten jakaminen ja varastointi, palaute	www-keskustelualue, puhelintuki	Perustamisvaihe, palveluvaihe

Tutkimuksen ja opetuksen tukipalveluihin sisältyy käyttäjien kaikenlainen avustaminen, kysymyksiin vastaaminen, käyttäjien vertaistuen mahdollistaminen ja opetukseen soveltuvien esimerkkiaineistojen laadinta.

Jotta aineistoista ja menetelmistä opitut asiat saadaan laajempaan käyttöön, muiden kuin arkaluontoisten tai teknisten tietojen lähetys ja vaihto tapahtuu MIDRAS-järjestelmässä keskustelu- ja tukifoorumin kautta. Foorumin käyttäjiä ovat tutkijat, viranomaisten edustajat ja MIDRAS-hallinnon edustajat. Foorumilla voi käydä vapaamuotoista keskustelua sekä lisäksi julkaista muiden käyttöön työkaluja, dokumentteja ynnä muuta sähköistä sisältöä. Foorumilla olevat viestit ja sisällöt ovat julkisia, ellei lähettäjä ole erikseen määrittänyt niitä rajoitetuiksi.

Foorumin kautta jaetaan muun muassa ylläpidolliset tiedotukset, hyväksi havaitut käytännöt, metatietojen parannusehdotukset, aineistojen korjauspyynnöt ja aineiston käsittelyssä apuna käytetyt menetelmät. Foorumi on muuten varsin tavallinen keskustelualue, mutta siinä olevat sisällöt, viestit ja kommentit voi luokitella liittyviksi tiettyihin aineistoihin, projekteihin, viranomaisiin ja niin edelleen. Lisäksi foorumi on integroitu tiedostojen lähetys- ja vastaanottopalveluun (katso edellinen osio) siten, että foorumiin on helppoa lähettää sisältöjä etätyöpöydältä ja foorumin sisältöjä etätyöpöydälle tarkastusmenettelyn kautta.

### 7.3. Toimintamallin prosessit ja toimijoiden vastuut

Tässä luvussa kuvataan MIDRAS-toimintamallin ja -järjestelmän keskeisimmät prosessit. Kuvaukseen sisältyvät toimijoiden vastuut eri tilanteissa ja MIDRAS-toimintamallin prosessien eteneminen. Selvityshankkeen aikana vertailtiin useita eri vaihtoehtoisia ratkaisuja eri prosessien parhaaksi toteutustavaksi ja menetelmiksi. Nämä vaihtoehdot ja vertailujen tulokset on esitelty tarkemmin liitteessä 2.

MIDRAS-toimintamallissa pitää määrittää selkeästi, kenen vastuulle kuuluu mikäkin tehtävä ja miten eri toimijoiden välinen yhteistoiminta on järjestetty. Vastuiden jakautumista suunniteltaessa on noudatettu seuraavia periaatteita:

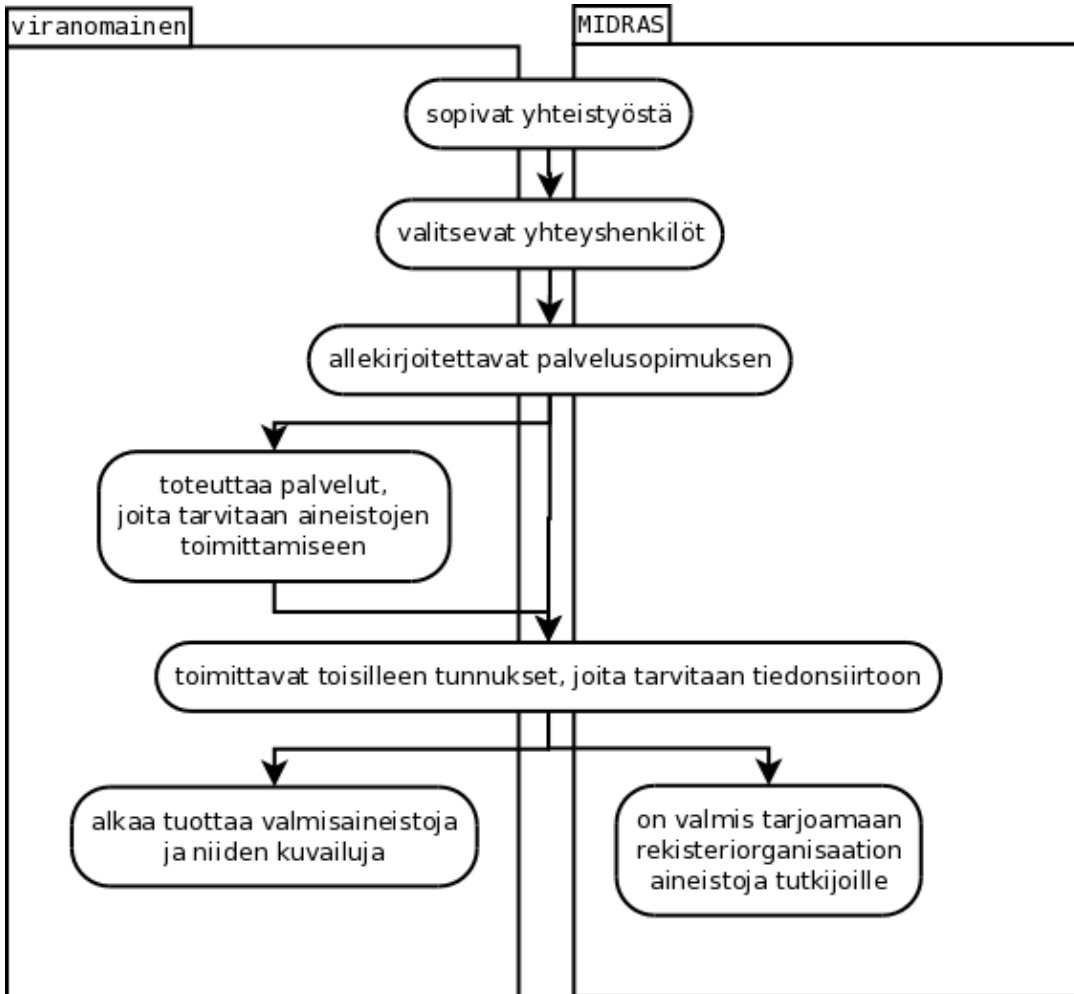
1. Tehtävät jaetaan karkeasti toimijoiden ydintehtävän mukaan.
2. Kehitystehtävät jaetaan tarvittavan osaamisen mukaan.
3. Lain asettamat velvoitteet eri osapuolille täyttyvät.

4. Osapuolet pystyvät useimmissa tapauksissa muuttamaan sisäistä toimintaprosessiaan tarvitsematta muuttaa yhteistyökäytäntöjä.

### 7.3.1. Viranomaisten liittyminen MIDRAS:iin

Kuvassa 5 on kuvattu, mitä toimia tarvitaan, kun uusi viranomainen haluaa alkaa toimittaa aineistojaan tutkimuskäyttöön MIDRAS-järjestelmän kautta.

Kuva 5 Viranomaisen liittyminen MIDRAS-järjestelmään



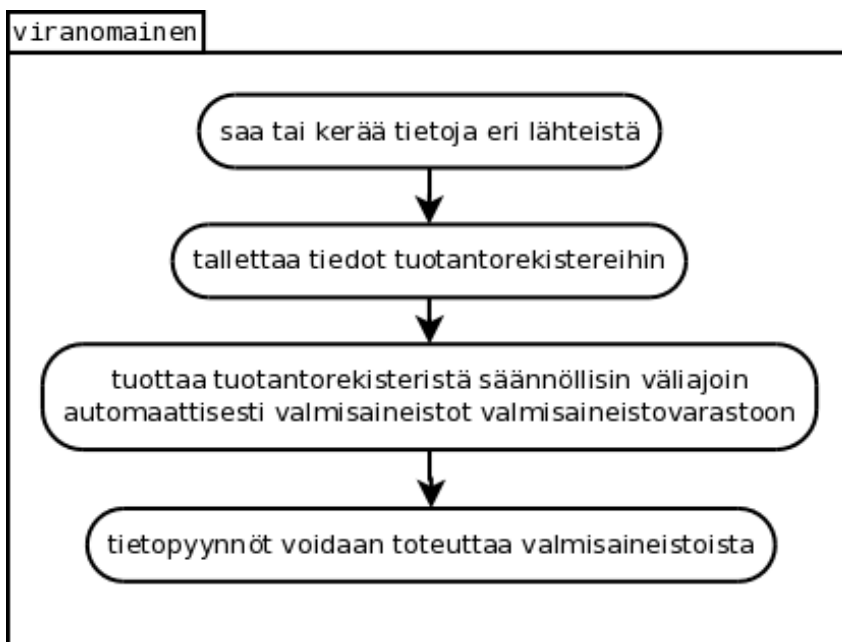
### 7.3.2. Tutkimusaineistojen tuottaminen

Tutkimusaineistojen muodostamisprosessi ei sinänsä kuulu MIDRAS-toimintamallin määritettäväksi, vaan eri viranomaisilla saattaa olla hyvinkin erilaisia tapoja ja prosesseja, miten rekisteriaineistot saadaan primäärikäytöstään tuotantorekistereinä sekundaariseen tutkimuskäyttöön. MIDRAS-toimintamalli antaa kuitenkin niin uudenlaiset välineet aineistojen toimittamiseen, että myös tutkimusaineistojen muodostamisprosessi voi muuttua osittain tai kokonaan.

Suosittamme, että aineistojen tuottaminen ja toimittaminen järjestetään jatkossa kolmitasoisena palveluna:

1. *Tuotanto-* eli *primäärirekistereissä* pidetään ajantasaista, muuttuvaa tuotantodataa, jota käytetään ensisijaiseen tarkoitukseensa.
2. *Valmisaineistovarastossa* pidetään tuotantorekisteristä automaattisesti ja säännöllisesti tuotettua, vakaampaa aineistoa, joka lisäksi saattaa olla kerätty sopivampaan muotoon tilastointia ja tutkimusta varten. Aineisto on muodossa, josta voidaan helposti poimimalla muodostaa yksittäisten tutkimusten aineistoja: kaikki yhteismitalliset tiedot (esimerkiksi eri vuosien tiedot, tai erilaiset aikajaksotiedot) on yhdistetty valmiiksi yhteen matriisiin.
3. Tiedot lähetetään MIDRAS-järjestelmään *tutkimusaineistopalvelimen* (push-malli) tai *aineistopalvelun* (pull-malli) kautta. Näistä palveluista saa vain pseudonymisoitua, tutkimusluvan mukaista aineistoa. Rekisteritiedot annetaan tutkijan käyttöön mahdollisimman alkuperäisessä muodossa ja aineiston tulkinta ja johdettujen muuttujien luonti jätetään tutkijan tehtäväksi.

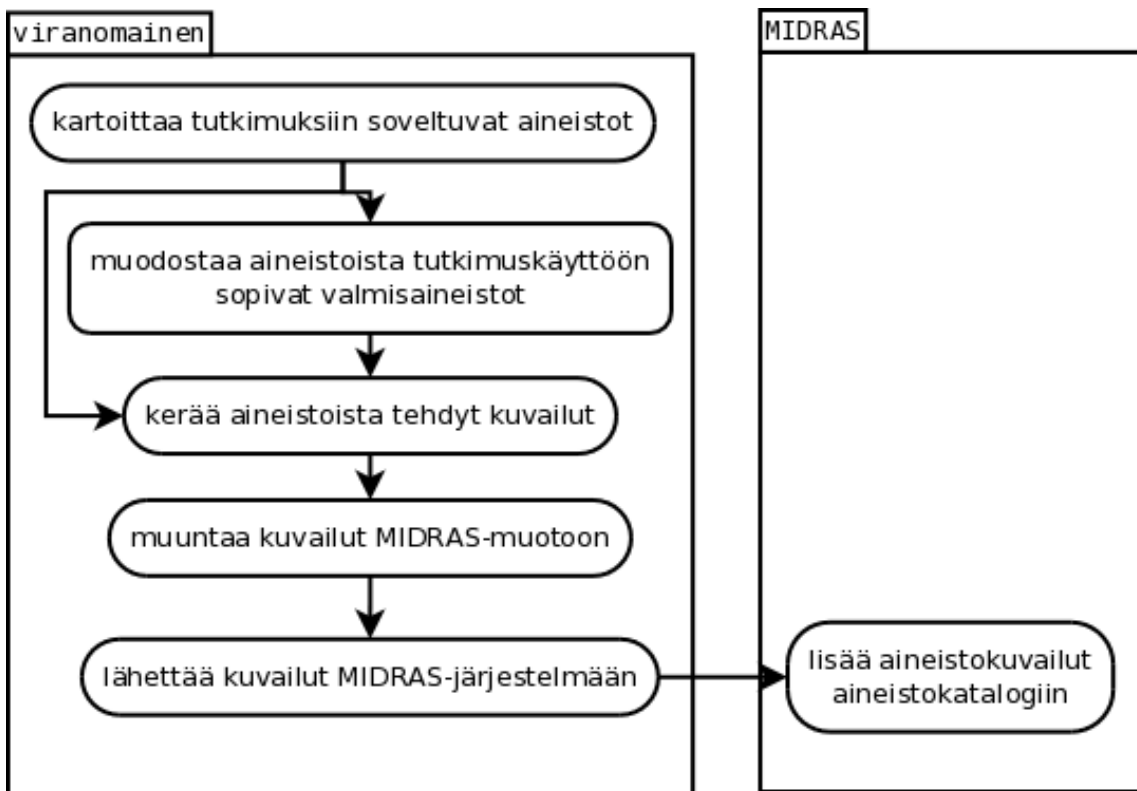
Kuva 6 valmisaineistovaraston ylläpito



### 7.3.3. Aineistojen kuvailu

Aineistojen tutkimuskäytön kannalta aineistojen kuvailu, eli aineistojen metatiedot, on keskeisessä asemassa. Aineistojen kuvailussa erityinen haaste on se, että korkealaatuisten kuvailujen tuottamiseen tarvitaan sekä sisältöasiantuntemusta yksittäisten metatietojen laadun takaamiseksi että monipuolista yhteistyötä ja vakiointia metatietojen yhteensopivuuden ja aineistokatalogin toimivuuden varmistamiseksi. Periaatteessa kaikki saatavilla oleva kuvailu parantaa aineistojen käytettävyyttä, mutta yhtenäiset kuvailukäytännöt parantavat huomattavasti metatietojen jatkokäytettävyyttä. Seuraavassa kaaviossa on kuvattu aineistojen kuvailun prosessi.

Kuva 7 Aineistojen kuvailu



Selvityshankkeen ohessa on tehty metatietostandardeista alustava ehdotus, joka on kuvattu tämän selvityksen liitteessä 6. Jotta aineistojen kuvailutyö on mahdollisimman helppoa saada alkuun, metatiedon vaatimusmäärittely on jaettu vaativuusluokkiin sen mukaan, kuinka keskeistä mikin metatieto on kuvailun tavoitteiden kannalta, ja huolehdittu siitä, että keskeisimpien tietojen antaminen ei ole kohtuuton työ.

Taulukko 7 Toimijoiden vastuut aineistojen kuvailussa

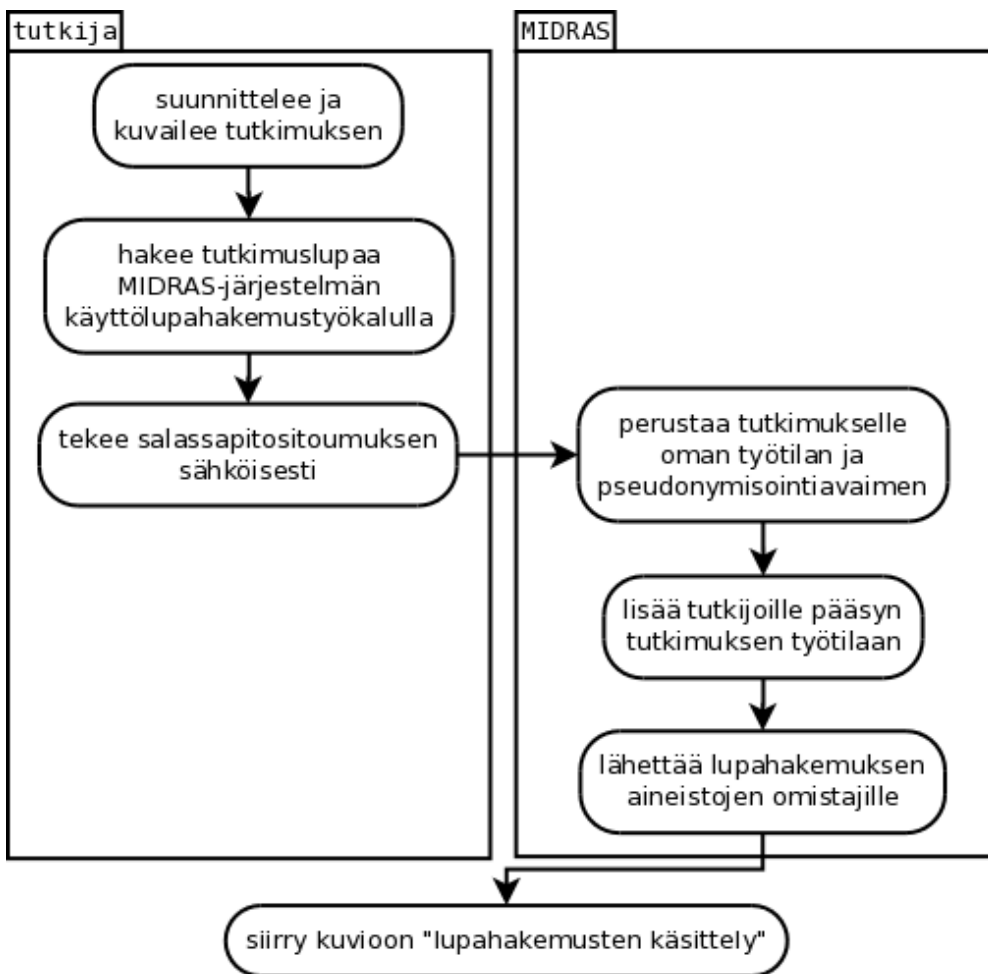
Aineistojen tuottajat vastaavat:	MIDRAS-hallinto vastaa:
<ul style="list-style-type: none"> <li>• Aineistojen kuvailu MIDRAS-järjestelmän ohjeiden mukaisesti</li> <li>• Metatietojen kuvailu (mm. julkisuus)</li> <li>• Metatietojen lähettäminen MIDRAS-järjestelmään</li> <li>• Metatietostandardien kehittämiseen osallistuminen</li> <li>• Metatietojen ylläpito ja päivitys</li> </ul>	<ul style="list-style-type: none"> <li>• Aineistojen kuvailun ohjeistus (metatietostandardit), työkalut ja käyttäjätuki</li> <li>• Metatietojen vastaanotto- ja säilöntäpalvelun ylläpito</li> <li>• Metatietojen hallinnointi kuvailijoiden vaatimusten mukaisesti</li> <li>• Metatietojen selailu- ja etsintäkäyttöliittymien ylläpito</li> <li>• Metatietostandardien kehittämisen koordinointi</li> </ul>

#### 7.3.4. Käyttöluvat

MIDRAS-järjestelmässä aineistojen käyttöluvat perustuvat edelleen tietotarpeelle, eli käyttöluvat myönnetään tutkimussuunnitelman perusteella. Aineistojen omistajat (viranomaiset) määrittävät toimintaansa ohjaavien lakien mukaisesti, mihin tai minkälaiseen aineistoon tutkimussuunnitelman perusteella annetaan

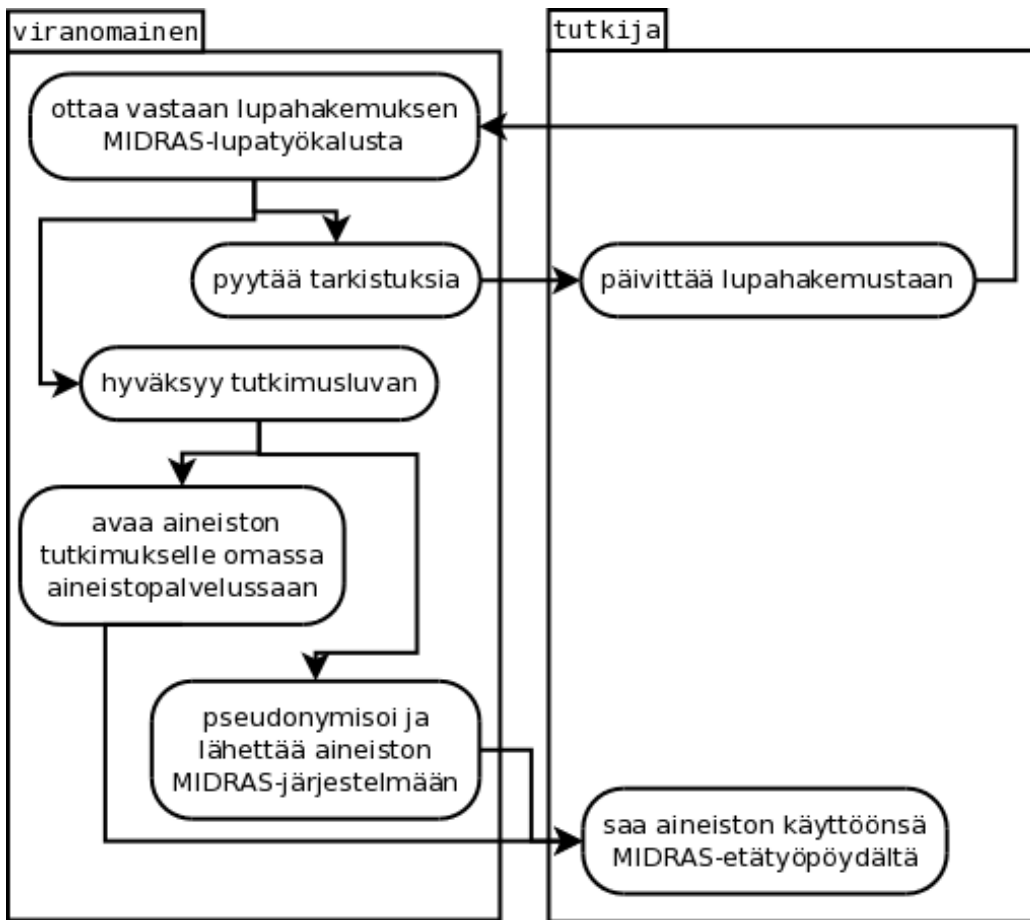
käyttölupaa. Kun tutkija hakee käyttölupaa, käyttölupahakemus lähetetään lupahakupalvelimen kautta sähköisesti edelleen niille viranomaisille, joiden tietoja tarvitaan tutkimuskohortin määrittämiseen ja varsinaisiin tutkimusaineistoihin, ja nämä hyväksyvät hakemuksen. Jos tutkija haluaa käyttää tutkimuksessaan tutkimusaineistoa, joka on muodostettu aiemmin toista tutkimusta varten ja säilötty MIDRAS-järjestelmään, tutkija hakee lupaa tutkimusaineiston jatkokäyttöön niiltä viranomaisilta, joiden toimittamien tietojen perusteella alkuperäinen tutkimusaineisto on muodostettu. Seuraavassa kaaviossa on kuvattu käyttölupaprosessi tutkijan näkökulmasta.

Kuva 8 Uuden tutkimuksen perustaminen



Seuraavassa kaaviossa on puolestaan kuvattu käyttölupaprosessi viranomaisen näkökulmasta.

Kuva 9 Lupahakemusten käsittely



Yllä kuvattu prosessi käydään läpi jokaisessa tutkimukseen aineistoja toimittavassa viranomaisessa joka kerran, kun tutkija tekee uuden tutkimusluvan tai päivittää vanhaa.

Taulukko 8 Toimijoiden vastuut lupamenettelyissä

Tutkija vastaa:	Aineiston omistajat (viranomaiset) vastaavat:	MIDRAS-hallintotaho vastaa:
<ul style="list-style-type: none"> <li>Tutkimuksen suunnittelu</li> <li>Tutkimuksen poiminnan ja aineistotarpeen suunnittelu ja kuvailu</li> <li>Tutkimusprojektiin osallistuvien tutkijoiden käyttäjätunnusten hakeminen MIDRAS-järjestelmään</li> <li>Käyttölupahakemuksen lähettäminen MIDRAS-lupahakupalvelulla</li> </ul>	<ul style="list-style-type: none"> <li>MIDRAS-lupahakupalvelun vaatimusten määrittäminen</li> <li>Käyttölupahakemusten sisällöllinen tarkastaminen ja käsittely</li> <li>Tutkimuksessa tarvittujen aineistojen määrittely ja tarpeen vaatiessa muodostaminen</li> <li>Käyttölupien myöntäminen</li> </ul>	<ul style="list-style-type: none"> <li>Tutkijoiden käyttäjätunnusten perustaminen</li> <li>Etätyöpöydän käytön seuranta</li> <li>Lupahakupalvelun ylläpito</li> <li>Käyttöoikeustyökalun ylläpito</li> <li>Käyttäjätuki</li> </ul>

### 7.3.5. Tutkimusaineistojen toimittaminen

Eri viranomaisten osalta MIDRAS-järjestelmässä noudatetaan eri toimintamalleja sen mukaan, millaisesta aineistosta on kysymys ja millaisia valmiuksia aineiston toimittavalla viranomaisella on toteuttaa rajapinta-pohjainen aineistopalvelu. Suosituksemme on, että MIDRAS-toimintamallissa jokaisella viranomaisella on kaksi vaihtoehtoa:

1. *MIDRAS-järjestelmän vastaanottopalvelu (ns. push-malli):*

MIDRAS-järjestelmä tarjoaa tietojärjestelmärajapinnan, jonka kautta viranomainen voi lähettää, päivittää ja poistaa tuottamansa aineistot sekä määrittää kenellä on niihin käyttöoikeus.

2. *Viranomaisen aineistopalvelu (ns. pull-malli):*

Viranomainen tarjoaa tietojärjestelmärajapinnan, jonka kautta MIDRAS-järjestelmä voi hakea aineistoja säännöllisesti tai silloin, kun tutkija pyytää niitä etätyöpöydän kautta. Tietopyynnön yhteydessä MIDRAS-järjestelmä lähettää tiedot siitä, mitä tietoja pyydetään kenen oikeuksilla, sekä tietopyynnön tunnistautumistiedot.

Mikäli käyttöluva on haettu aiemmin muodostetun tutkimusaineiston jatkokäyttöön, tutkimusaineisto avataan yksinkertaisesti MIDRAS-järjestelmän tutkimussäilöstä uusille tutkijoille.

Taulukko 9 Yhteenveto vastuista

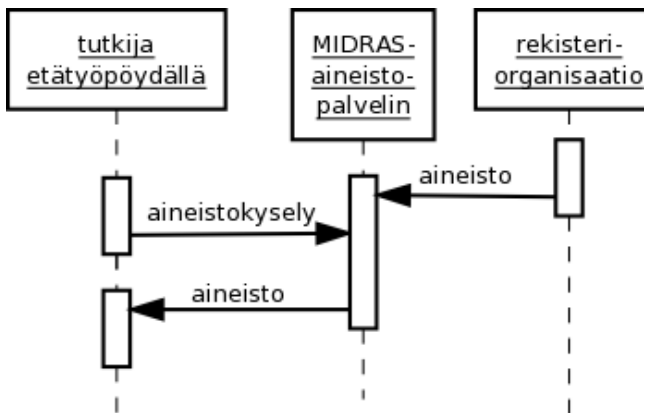
Aineiston omistaja:	MIDRAS-hallinto:
<ul style="list-style-type: none"><li>• Aineiston sisältö ja pseudonymisointi</li><li>• Käyttöoikeuksien määrittäminen</li><li>• Aineiston lähetys (push-mallissa)</li><li>• Aineistopalvelun ylläpito ja seuranta (pull-mallissa)</li><li>• Tietopyyntöjen auktorisointi (pull-mallissa)</li><li>• Aineistojen tarjoaminen aineistopalvelun kautta (pull-mallissa)</li></ul>	<ul style="list-style-type: none"><li>• Aineiston luovuttaminen aineiston omistajan määrittämien tutkimusprojektien / tutkijoiden käyttöön käyttöluvan keston ajaksi</li><li>• Aineistojen vastaanottopalvelun ylläpito</li><li>• Aineistojen eheyden tarkistus metatietojen perusteella lähetysvirheiden havaitsemiseksi</li><li>• Aineiston käyttökopion säilytys tarvittaessa</li><li>• Aineiston poistaminen MIDRAS-järjestelmästä aineistoa käyttäneen tutkimusprojektin loppuessa</li><li>• Tietopyynnön auktorisointitietojen välittäminen aineistopalvelulle (pull-mallissa)</li><li>• Aineistojen käytön seuranta ja pääsynhallinta käyttöoikeuksien mukaisesti</li></ul>

#### *Aineistojen toimittamisen push-malli*

Push-mallissa viranomainen avaa aina halutessaan tietoyhteyden MIDRAS-järjestelmään lisätäkseen uusia aineistoja, päivittääkseen jo lähetettyjä tai poistaakseen vanhoja. Tutkimuksessa käytettyä aineiston kopio-

ta säilytetään väistämättä MIDRAS-järjestelmässä, joten aineiston käytön seuranta ja tekninen pääsynhallinta ovat MIDRAS-ylläpidon vastuulla.

Kuva 10 Aineiston kulku push-mallissa

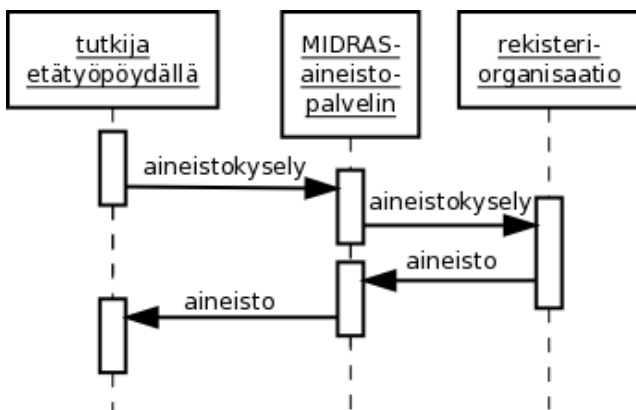


Tietojen poiminnan ja pseudonymisoinnin tutkimusta varten voi myös hoitaa automaattilla tai esim. tietokantanäkymänä siten, ettei aineistosta ole tarpeen pitää erikseen tallessa useampaa, eri tavoin poimittua versiota. Mikäli eri tutkimusten poiminnat tuotetaan samasta pohja-aineistosta automaattisesti, on mahdollista pohja-aineiston päivittyessä tuottaa automaattisesti myös kaikille tutkimuksille päivitettyt versiot.

### *Aineistojen toimittamisen pull-malli*

Pull-mallissa MIDRAS-järjestelmä avaa tietoyhteyden viranomaisen järjestelmään tehdäkseen tietopyyntöjä tai välittääkseen tutkijoiden tietopyyntöjä. Aineistoa hakiessaan MIDRAS-järjestelmä todistaa tietopyynnön autenttisuuden ja antaa viranomaisen aineistopalvelulle tiedot, mihin tutkimusprojektiin, kenen oikeuksilla ja mitä tietoja halutaan.

Kuva 11 Aineiston kulku pull-mallissa



### *Push- ja pull-mallin vertailua*

Push-mallin etuna on se, että se tarjoaa aineiston tuottajalle paljon vapautta aineistolähetysten järjestämisessä: aineistojen tuottamisen ja lähetyksen voi automatisoida tai sen voi tehdä kertatyönä aina tarpeen vaatiessa. Lisäksi push-malli ei edellytä viranomaiselta aineistopalvelun ylläpitoa.

Pull-mallin etuna on se, että aineistojen käytön läpinäkyvyys viranomaisen suuntaan lisääntyy, eikä aineiston kopiota tarvitse säilöä pitkiä aikoja MIDRAS-järjestelmässä, koska sen voi tarvittaessa hakea uudestaan aineistopalvelusta. Lisäksi aineistopalvelua voidaan mahdollisesti käyttää muunkinlaiseen tietojenvälitykseen kuin MIDRAS-järjestelmän edellyttämään tutkimusaineistojen välitykseen. Vain pull-mallissa on mahdollista, että aineistojen käyttäjä (tutkija) määrittää yhdestä aineistosta tarvitsemansa tiedot toisen aineiston perusteella ja saa yhdistellyn kyselyn tulokset itselleen välittömästi.

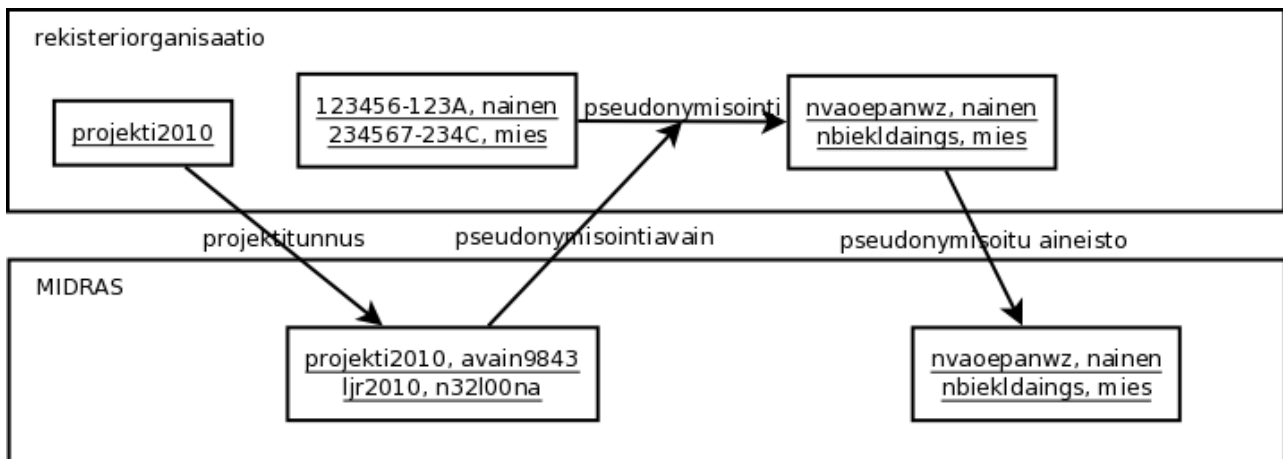
Suosittamme, että viranomaiset siirtyvät vähitellen noudattamaan pull-mallin mukaista toimintatapaa.

#### **7.3.6. Pseudonymisointi ja aineistojen yhdistely**

MIDRAS-järjestelmän lähtökohta on, ettei aineistoja pitäisi tarjota järjestelmään henkilötunnuksellisinä. Toisaalta useammista lähteistä tulevien tietojen yhdistely edellyttää jonkinlaisia henkilön tai muun yksikön yksilöiviä tunnisteita aineistoissa. Niinpä aineistot pseudonymisoidaan yhdistelyä varten jo ennen MIDRAS-järjestelmään tarjoamista. Pseudonymisointi tarkoittaa tunnisteiden korvaamista toisilla, vain tutkimusprojektin aineistossa käytetyillä tunnisteilla.

Koska aineistot on tarkoitus lähettää MIDRAS-järjestelmään valmiiksi pseudonymisoituina, MIDRAS määrittää kaikille aineiston tuottajille yhteisen tavan pseudonymisoida aineistoja. Pseudonyymien projektikohtaisuus varmistetaan käyttämällä pseudonymisoinnissa projektikohtaista salalauseetta, joka on vain niiden aineiston tuottajien tiedossa, jotka toimittavat aineistoja kyseiseen tutkimusprojektiin. Tämä salalause voidaan antaa kaikkien aineiston tuottajien tietoon MIDRAS-järjestelmän aineistonvastaanottopalvelun kautta, tai aineistojen tuottajat voivat sopia sen keskenään, mikäli katsotaan tietosuojasyistä hyväksi olla viemättä salalauseetta ollenkaan MIDRAS-järjestelmään.

Kuva 12 Pseudonymisointi MIDRAS-järjestelmässä



Kun pseudonymisoidut aineistot ovat MIDRAS-järjestelmässä, MIDRAS pseudonymisoi aineiston vielä toiseen kertaan, jotta henkilö, joka on kaksoisasemassa aineiston tuottajana ja käyttäjänä, ei pystyisi avaamaan pseudonyymeja. Kaksinkertaisesti pseudonymisoitu aineisto annetaan tutkijan käyttöön MIDRAS-järjestelmässä, ja tutkija yhdistelee aineistojen tiedot pseudonyymien perusteella.

Taulukko 10 Yhteenvetotaulukko toimijoiden vastuista aineistojen pseudonymisoinnissa ja yhdistelyssä

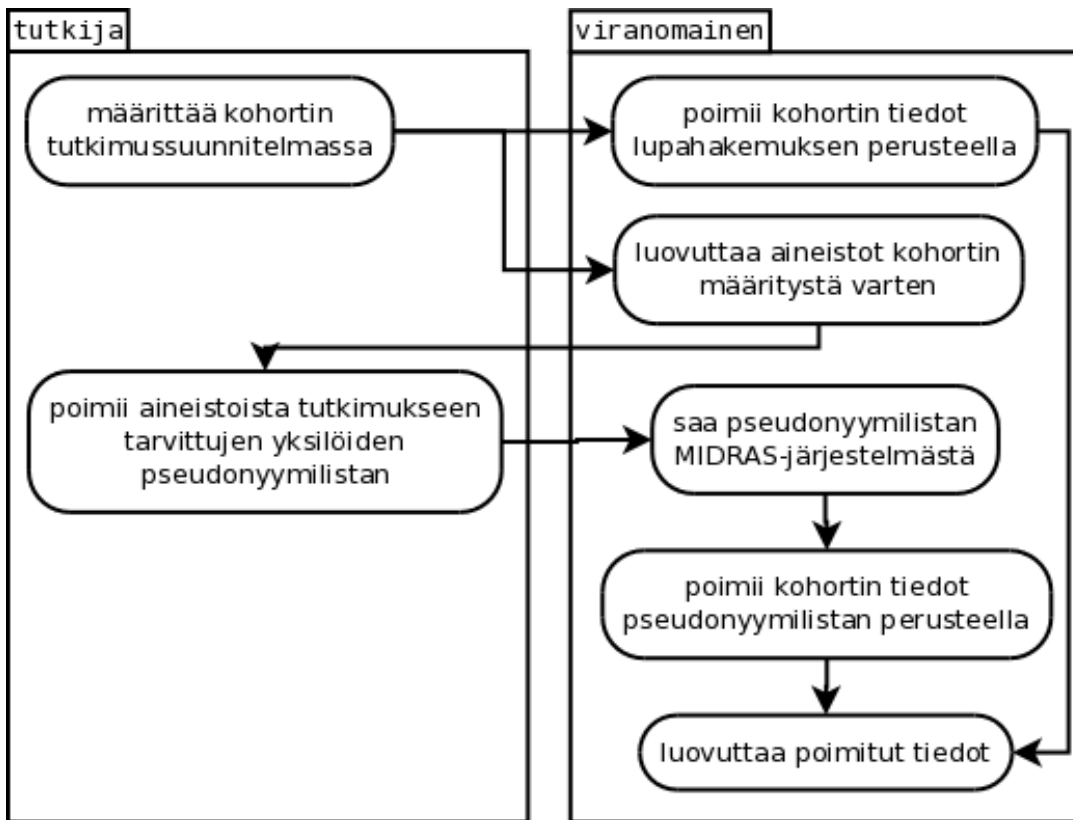
Aineiston tuottaja vastaa	MIDRAS-hallinto vastaa	Tutkija vastaa
<ul style="list-style-type: none"> <li>Pseudonymisointialalauseen hakeminen MIDRAS-järjestelmästä tai sopiminen muiden aineiston tuottajien kanssa</li> <li>Aineiston pseudonymisointi</li> <li>Aineiston poiminta MIDRAS-järjestelmästä saadun pseudonymilistan perusteella</li> <li>Pseudonymisoidun aineiston tarjoaminen MIDRAS-järjestelmään</li> </ul>	<ul style="list-style-type: none"> <li>Pseudonymisoinnin ohjeistus ja työkalut</li> <li>Pseudonymisointialalauseen tiedottaminen tutkimusprojektin aineiston tuottajille</li> <li>Kohortin tiedottaminen pseudonymilistana poimintaa varten</li> <li>Tutkijoiden omien aineistojen tarkastus, pseudonymisointi ja lisääminen MIDRAS-järjestelmään</li> </ul>	<ul style="list-style-type: none"> <li>Tutkimuskohortin muodostaminen poiminta-aineistojen perusteella</li> <li>Aineistojen yhdistely pseudonyymien perusteella</li> <li>Mahdollisesta pseudonyymien avaamistarpeesta tiedottaminen</li> <li>Omien aineistojen tuottaminen, suostumusten kerääminen tutkimusjoukolta tietojen yhdistelemiseksi rekisteritietoihin ja aineiston lähettäminen MIDRAS-järjestelmään pseudonymisoitavaksi</li> </ul>

### *Tutkimuskohortin määrittäminen monimuuttajaisissa aineistopoiminnissa*

Useissa tutkimusprojekteissa tutkimuskohortti muodostetaan ensin erillisen "poiminta-aineiston" sisältämien tietojen perusteella, ja varsinaisten tutkimusaineistojen yksilöt poimitaan tämän kohortin yksilölistan perusteella. MIDRAS-suositus on, että tutkimuskohortin muodostaa tutkija, mutta on myös mahdollista, että jokin viranomaislainen muodostaa tutkimuskohortin tutkimussuunnitelmassa olevan kuvauksen perusteella.

la. Kohortin pseudonyymilista annetaan aineiston tuottajalle tietopyynnön yhteydessä (pull-mallissa) tai MIDRAS-järjestelmän aineistonvastaanottopalvelun kautta (push-mallissa).<sup>3</sup>

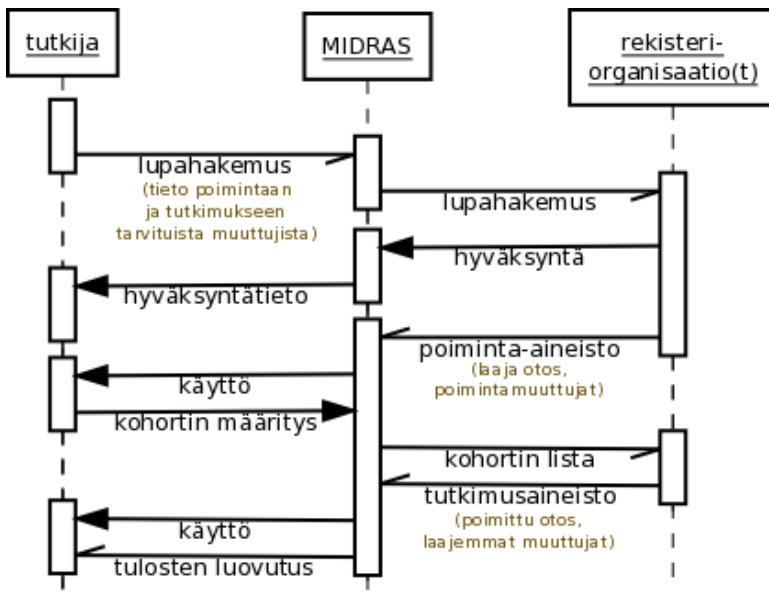
Kuva 13 Kohortin muodostaminen ja aineiston poiminta



Kuitenkin, jos tutkimuksessa on mukana tilastolain alaisia aineistoja, kaikki aineistot täytyy tällä hetkellä lähettää tilastoviranomaiselle yhdisteltäväksi ja anonymisoitavaksi, jotta tilastolain mukainen anonymisointivelvoite täyttyy. Tällöin tilastoviranomainen toimittaa aineistot MIDRAS-järjestelmään, ikään kuin olisi kaikkien aineistojen tuottaja.

<sup>3</sup> Push- ja pull-mallit on selitetty tarkemmin kohdassa "Tutkimusaineistojen toimittaminen" (6.3.5).

Kuva 14 Kohorttien välittämisen kommunikaatioprosessi



### Tutkijan omien aineistojen yhdistely

Joissakin tutkimusprojekteissa tutkijalla on oma aineisto, jonka kohteilta on erikseen saatu suostumus tietojen yhdistämiseksi rekisteritietoihin. Tällaisessa tilanteessa aineisto lähetetään tarkistettavaksi MIDRAS-ylläpidolle, joka lisää sen pseudonymisoituna MIDRAS-järjestelmään samaan tapaan, kuin jos aineisto olisi tutkijasta riippumattoman aineiston tuottajan tarjoama.

### 7.3.7. Käyttäjien ja palveluiden tunnistautuminen

Lähes kaikki MIDRAS-palvelut edellyttävät jonkinlaista tunnistautumista. Palvelut jakautuvat tunnistautumisen kannalta neljään luokkaan:

- tietosuojan kannalta kriittiset käyttäjien palvelut, joissa pääsee käsiksi arkaluontoisiin aineistoihin.
- muut käyttäjäpalvelut, jotka ovat vähemmän kriittiset tietosuojan kannalta.
- palveluiden väliset, automatisoitaviin prosesseihin suunnitellut rajapinnat.
- avoimet palvelut, jotka eivät vaadi minkäänlaista tunnistautumista.

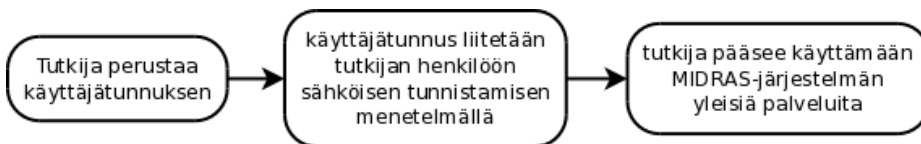
Ensimmäisen luokan palveluihin kuuluvat etätyöpöytä ja sillä olevat aineistot. Näiden palveluiden käyttö edellyttää vahvaa tunnistusta. Vahva tunnistus toteutetaan yhdistämällä käyttäjätunnus-salasanapari kertakäyttösalasanaan, joka lähetetään tekstiviestitse käyttäjän matkapuhelimeen kirjauduttaessa sisään.

Toisen luokan palveluihin kuuluvat lupahakemuspalvelu, etätyöpöydän käyttäjien tukifoorumi, käyttöoikeustyökalu, tulostenluovutuspalvelu, tutkijan omien aineistojen vastaanottopalvelu, aineistojen käytön seuranta työkalu ja metatietojen muokkaustyökalu. Näihin palveluihin tunnistaudutaan samalla käyttäjätunnus-salasanaparilla kuin ensimmäisen luokan palveluihin, mutta käyttö ei edellytä vahvaa tunnistusta eikä matkapuhelimeen lähetettävää kertakäyttösalasanaa käytetä.

Kolmannen luokan palveluita ovat aineistojen vastaanottopalvelu, viranomaisen aineistopalvelut, metatietojen vastaanottopalvelu ja MIDRAS-järjestelmän sisäiset tietoyhteyspalvelut kuten etätyöpöydän ja aineistotietokannan välinen yhteys. Näissä palveluissa palvelin ja asiakas tunnistautuvat molemminpuolisesti kryptografisilla varmenteilla.

Neljännän luokan palveluita ovat etätyöpöydän käyttäjien tukifoorumin arkisto, aineistokatalogi ja siellä olevat metatiedot julkiselta osaltaan, ohjeet, käyttäjätuki ja järjestelmän tiedotteet. Nämä www-palvelut eivät vaadi käyttäjän tunnistamista ja niiden tietosisällöt saavat päätyä hakukoneisiin. Tietojen julkisuus lisää myös yleistä tietoisuutta tehdystä ja tekeillä olevasta tutkimuksesta, tutkimusmenetelmistä, rekisteriaineistojen sisällöstä ja MIDRAS-järjestelmästä.

**Kuva 15 Käyttäjätunnuksen luominen**



Mahdollisia sähköisen tunnistamisen menetelmiä ovat esimerkiksi HAKA-tunnistusfедераatio, Virtutunnistusfедераatio, VETUMA-tunnistus, HST-sirukortti.

**Kuva 16 Käyttäjätunnuksen poisto**

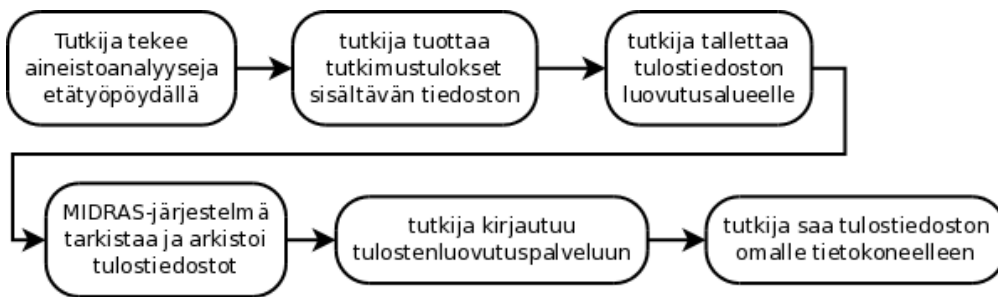


### 7.3.8. Aineistojen käsittely ja tutkimustulosten luovuttaminen

Aineistojen käsittelyä varten MIDRAS-järjestelmässä on etätyöpöytä, jolle tutkija kirjautuu ja jossa tutkija tekee varsinaisen aineistojen käsittelytyön. Tulokset saa omaan haltuun tulostenluovutuspalvelun kautta. Tutkija vastaa siitä, että tuloksissa ei ole tietosuojaa vaarantavia tietoja. MIDRAS-hallinto huolehtii tulosten tarkastuksesta, seurannasta ja tallentamisesta jatkokäyttöä varten.

Seuraavissa kaavioissa on kuvattu tutkimustulosten luovutusprosessi sekä tutkimuksen päättäminen.

**Kuva 17 Tutkimustulosten luovutus.**



Kuva 18 Tutkimuksen loppuminen.



MIDRAS-käyttösovelmus velvoittaa tutkijan jakamaan käyttämänsä työskentelymenetelmät (myös itse tehdyt ohjelmat ja automaattioskriptit) sekä tarjoamaan tutkimusaineistot jatkotutkimuksen käyttöön kahden vuoden kuluttua tutkimuksesta. Näiden työskentelymenetelmien dokumentointi ja aineistojen kuvailu on tutkijan vastuulla. Aineiston tuottajat voivat esittää MIDRAS-ylläpidolle vaatimuksia ja kehittää työkaluja tarkistusmenettelyiden suhteen.

Taulukko 11 Toimijoiden vastuut aineistojen käsittelyssä

Tutkija vastaa:	MIDRAS-hallinto vastaa:
<ul style="list-style-type: none"> <li>Aineistojen käsittely, analyysi ja tulkinta etätyöpöydällä</li> <li>Tutkimustulosten muodostaminen ja siirtäminen luovutusalueelle</li> <li>Tutkimuksessa muodostettujen aineistojen ja työvälineiden kuvailu jatkokäyttöä varten</li> <li>Tietosuojan, tutkimusetiikan ja muiden velvoitteiden ottaminen huomioon aineistoja käsiteltäessä ja tuloksia kirjoitettaessa</li> <li>Tarvitsemiensa työkalujen määrittely ja työkaluvalikoiman laajennus yhteistyössä MIDRAS-ylläpidon kanssa</li> <li>Henkilörekisterin lakitekkinen pitovastuu (henkilöaineistoja käsiteltäessä) ja rekisteriseloste-ehdotuksen tarkistus</li> </ul>	<ul style="list-style-type: none"> <li>Etätyöpöydän ylläpito ja tietoturva</li> <li>Etätyöpöydän työkaluvalikoiman ylläpito</li> <li>Tutkijoiden omien työkalujen tarkastus ennen etätyöpöydälle viemistä</li> <li>Etätyöpöydän käyttäjätuki siltä osin, kuin se on vain MIDRAS-järjestelmää koskevaa eikä ohjelmien yleiskäyttötukea</li> <li>Tulostenluovutuspalvelun ylläpito ja seuranta</li> <li>Henkilörekisteriseloste-ehdotuksen automaattinen muodostus</li> </ul>

## 7.4. MIDRAS-järjestelmän tekninen toteutus

Tässä osassa kuvataan MIDRAS-järjestelmän palveluita ja niiden toteutusvaihtoehtoja yksityiskohtaisemalla tasolla. Tarkoituksena on kuvata ehdotettu MIDRAS-järjestelmä niin tarkasti, että sen pystyy kuvauksen perusteella toteuttamaan. Osio sisältää paljon teknisiä yksityiskohtia, joiden ymmärtäminen ei ole välttämätöntä MIDRAS-kokonaisuuden hahmottamisen kannalta. Tekniset ratkaisut kehittyvät koko ajan, eikä kaikkien yksityiskohtien lyöminen lukkoon etukäteen ole tarkoituksenmukaista. MIDRAS-järjestelmää on tarkoitus kehittää sitä mukaa, kuin helppokäyttöisempiä, tehokkaampia, monipuolisempia tai muuten parempia teknisiä ratkaisuja tulee tarjolle. Tässä osiossa määritellyt ratkaisut arvioidaan pysyviksi ainakin noin viiden vuoden aikajänteellä.

### 7.4.1. Järjestelmän palvelinrakenne

MIDRAS-järjestelmän tekninen toteutus koostuu eristetystä palvelinympäristöstä, jossa on neljänlaisia palvelimia:

#### 1. *etätyöpöytäpalvelimet*

Etätyöpöytäyhteyksiä tarjoavat palvelimet, jolla tutkijoiden työskentely-ympäristö toimii. Analyysi- ja tilasto-ohjelmat ajetaan etätyöpöytäpalvelimilla. Eri tutkimusprojektit on eriytetty toisistaan niin, että jokainen tutkimusprojekti näkee oman, erillisen virtuaalipalvelimensa. Etätyöpöytäpalvelimia voi olla useammanlaisia sen mukaan, millaiset tietojenkäsittelyvaatimukset tutkijoilla on: ainakin ajantasaisia Windows- ja Linux-ympäristöjä tarjotaan työskentely-ympäristöiksi.

#### 2. *aineistopalvelin*

Tietokantapalvelin, jonka kautta tutkijat saavat tutkimusaineistot käyttöönsä. Aineistopalvelin ei tarjoa muita palveluita kuin tietokantarajapinnan, jonka kautta tutkijat voivat kysellä aineistojen tietoja etätyöpöydältä.

#### 3. *oheispalveluiden palvelimet*

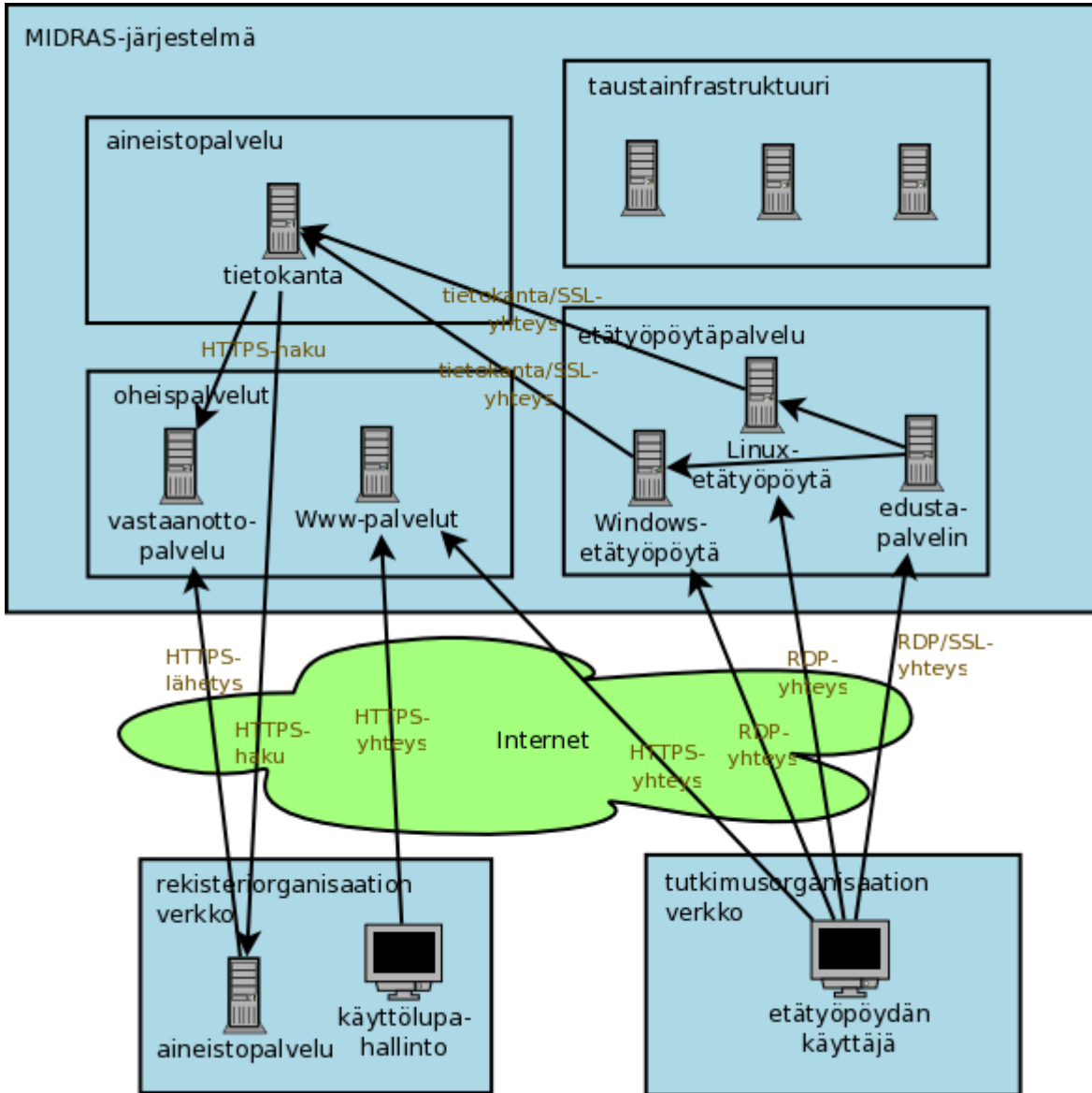
Www-palvelimia, joilla kaikki MIDRAS-järjestelmän oheispalvelut toteutetaan. Oheispalveluiden palvelimet käyttävät samaa käyttäjätunnistusta kuin muu MIDRAS-järjestelmä ja automaattit siirtävät tietoja muusta MIDRAS-järjestelmästä oheispalveluiden käyttöön, mutta muuten nämä palvelimet ovat erillisiä muusta MIDRAS-järjestelmästä.

#### 4. *infrastruktuuripalvelimet*

Muiden palvelimien toiminnan kannalta olennaiset palvelimet, kuten etätyöpöydän edustapalvelin, käyttäjätunnistuspalvelimet, päivityspalvelimet, varmuuskopiopalvelimet, verkkon seurantapalvelin, tietojärjestelmäseurantapalvelin, tietomurronhavaitsemispalvelin (IDS), lo-

kipalvelin, nimipalvelimet, aikapalvelimet, virtuaalipalvelimien hallinnointipalvelimet, palomuurit ja verkonhallintapalvelimet.

Kuva 19 Palvelinympäristö



#### 7.4.2. Tutkijan etätyöpöytä

Tutkijan etätyöpöytä toteutetaan tutkimusprojektikohtaisella virtuaalipalvelimella, jolle pääsee kirjautumaan käyttäen RDP-verkkoprotokollaa. Sekä Linux- että Windows-palvelimille on olemassa RDP-palvelintoteutukset: Windowsissa se on käyttöjärjestelmän mukana tuleva Remote Desktop Services (RDS) -palvelu, Linuxissa xrdp-palvelinohjelmisto. Leikepöydän käyttö, tiedostojen siirto ja muu tiedonsiirto RDP-

yhteyden kautta estetään palvelimen asetuksissa. Palvelimen asetuksissa edellytetään myös RDP-yhteyksiltä vahvaa salakirjoitusta.

Monien organisaatioiden palomuurissa on estetty RDP-yhteyksien muodostaminen organisaation verkon ulkopuolelle. Tästä syystä MIDRAS-järjestelmässä on myös etätyöpöydän edustapalvelin, joka ottaa vastaan SSL-tunneloituja RDP-yhteyksiä. Edustapalvelimella pidetään sekä Windowsin TS Gateway -palvelua että stunnel-palvelua. Ensin mainitun kautta etätyöpöydälle voi muodostaa RDP 6 -yhteyksiä, jotka näyttävät palomuurien kannalta https-yhteyksiltä, ja jälkimmäisen kautta etätyöpöydälle voi muodostaa SSL-käärittyjä RDP-yhteyksiä.

RDP-protokollaa käytetään Windows-työasemalta Remote Desktop Connection -ohjelmalla ja Linux- tai Mac-työasemalta rdesktop-ohjelmalla. Jos Windows-työasema on vanhempi kuin Windows 2003, Windows Vista tai Windows XP SP2, riittävä yhteyksien salakirjoitustaso vaatii yhteyksien erillistä tunnelointia SSL-tunnelin läpi.

#### **7.4.3. Tutkimusten eristäminen toisistaan**

Tutkimusaineistojen tietosuojan kannalta on olennaista, ettei eri tutkimusten tietoaineistoja pysty yhdistelemään keskenään. Osaltaan yhdistely on estetty sillä, että eri tutkimusprojektien aineistoissa yksilöillä tai muilla yksiköillä on eri pseudotunnisteet, mutta tämä yksinään ei riitä estämään aineistojen yhdistelyä välilisten tunnisteiden avulla. Tämän vuoksi eri joka tutkimusprojektilla on oma, virtuaalipalvelimella toteutettu etätyöpöytänsä, eli tutkimusprojektien etätyöpöydät on eristetty toisistaan niin kuin ne on eristetty muusta verkosta.

Jokaisen tutkimuksen etätyöpöydällä on oma osoitteensa. Tutkija määrittää, mitä tutkimusprojektia hän työstää, ottamalla etätyöpöytäyhteyden sen palvelimen osoitteeseen, jolla kyseisen tutkimusprojektin etätyöpöytä on. Eri etätyöpöydille kirjaudutaan samoilla käyttäjätunnuksilla; vain palvelimen osoite vaihtuu tutkimusprojektin mukaan.

Etätyöpöydältä käsin tutkijat hakevat tutkimusprojektinsa aineiston MIDRAS-järjestelmän aineistopalvelimelta. Aineistopalvelimelle kirjaudutaan varmenteella, joka on yksilöllinen kullekin tutkimusprojektille. Aineistojen hakuun käytetty varmenne talletetaan tutkimusprojektin etätyöpöydälle. Näin tutkimusprojektin tutkijat pystyvät hakemaan aineistopalvelimelta ne ja vain ne aineistot, jotka kuuluvat tutkimusprojektille.

#### **7.4.4. Aineistopalvelin**

Aineistopalvelimella toimii tietokantapalvelu, jonka kautta tutkijat pystyvät hakemaan tietoja tutkimusaineistosta. Lisäksi aineistopalvelimella ajetaan automaatteja, jotka hakevat tietoja aineistotietokantaan viiranomaisten aineistopalveluista ja MIDRAS-järjestelmän aineistonvastaanottopalvelusta.

Palvelussa käytetyltä tietokantaohjelmistolta ei edellytetä muuta, kuin että se tukee käyttäjävarmenteella tunnistettuja SSL-tietokantayhteyksiä. Esimerkiksi ilmainen MySQL-tietokantaohjelmisto täyttää tämän vaatimuksen. Tulevaisuudessa virtuaalitietokantojen teknologia saattaa kypsyä, jolloin tulee aiheelliseksi harkita virtuaalitietokantojen käyttöä. Virtuaalitietokantaratkaisussa on mahdollista hakea tietokantaan tiedot viranomaisen aineistopalvelusta vasta, kun käyttäjän tekemä kysely edellyttää tietojen hakua.

#### 7.4.5. Oheispalvelut

Kaikki MIDRAS-järjestelmän oheispalvelut ovat www-sovelluksia, joita käytetään salakirjoitettujen https-yhteyksien kautta. Palveluiden toteuttamiseen voidaan käyttää mitä tahansa www-sovellusten kehitysalustaa, esimerkiksi Apache + Python. Useat sovellukset vaativat myös taustalle tietojen tallennusjärjestelmää, jollaiseksi käy esimerkiksi MySQL-ohjelmisto.

#### 7.4.6. Tietoliikenneyhteydet

MIDRAS-järjestelmässä tutkimusten etätyöpöytäpalvelimet on eristetty muusta maailmasta palomuureilla. Erityisen tärkeää on, että etätyöpöytäpalvelimilta ei voi muodostaa ulos muita yhteyksiä kuin tietokantayhteydet aineistopalveluun.

Kaikki järjestelmän tietoliikenneyhteydet ovat salakirjoitettuja, paitsi tietyt yhteydet infrastruktuuripalvelimiin kuten nimipalvelimille ja päivityspalvelimille. RDP-, SSH- ja SSL-yhteyksissä (joihin myös https-yhteydet kuuluvat) käytetään salakirjoitusmenetelmiä, joissa avain on vähintään 128 bittiä pitkä, ja palvelin- ja asiakasvarmenteissa käytetään avaimia, jotka ovat vähintään 1024 bittiä pitkiä.

#### 7.4.7. Liittymät ja rajapinnat

Tietojen federaatio MIDRAS-järjestelmässä edellyttää monia rajapintoja:

- MIDRAS-järjestelmän aineistonvastaanottopalvelu tarjoaa viranomaisille rajapinnan aineistojen lähettämiseen, pseudonymisointisalalauseen hakemiseen ja tutkimuksen kohortin hakemiseen.
- viranomaisen aineistopalvelu tarjoaa MIDRAS-järjestelmälle (sekä mahdollisesti muille viranomaisen aineistoja hyödyntäville järjestelmille) rajapinnan aineistopyyntöjen tekoon.
- MIDRAS-järjestelmän metadatanvastaanottopalvelu tarjoaa viranomaisille rajapinnan aineistojen kuvailujen lähettämiseen.
- MIDRAS-järjestelmän metadatanreplikointipalvelu tarjoaa muille aineistokatalogeille rajapinnan metatietojen hakemiseen ja lähettämiseen.

Kaikki rajapinnat ovat HTTPS-pohjaisia, SSL-varmenteella autentikoituja ReST -palveluita. Palveluiden määrittelyissä noudatetaan Julkishallinnon perustietovarantojen rajapinnat -työryhmän (Valtiovarainministeriö, 2010a) suosituksia.

MIDRAS-järjestelmään otetaan vastaan aineistojen metatietoja samalla rajapinnalla kuin aineistojakin, ja lisäksi MIDRAS-järjestelmässä on rajapinta aineistojen kuvailujen vastaanottamiseksi aineistokatalogiin. Lisäksi MIDRAS-järjestelmään voidaan rakentaa muita rajapintoja metatietokannan integroimiseksi muihin järjestelmiin, esimerkiksi OAI-PMH-metatietorajapinta ja CKAN-tietovarastorajapinta.

#### 7.4.8. Tiedostomuodot

MIDRAS-toimintamallissa lähetetään monenlaisia tietoja tietojärjestelmien välillä. Niinpä pitää vakiodia, missä muodoissa tiedot lähetetään. MIDRAS-järjestelmän pitämiseksi yksinkertaisena jokaiselle tietotyypille on määritetty yksi tiedostomuoto. Tiedostomuodon valinnassa on pidetty mielessä seuraavia tavoitteita:

- tiedostomuoto riittää kuljettamaan kaiken informaation, jota sillä pitää siirtää.
- tiedostomuoto on avoin ja dokumentoitu.
- tiedostomuodon tuottamiseen ja käsittelyyn on valmiita työkaluja eri ympäristöissä.

Alla ovat erilaiset tietotyypit niille suositeltuihin tiedostomuotoineen.

- aineistot lähetetään UTF-8-koodattuina, otsakkeellisina sarkainerotettuina CSV-tiedostoina.
- aineistojen kuvailut, sekä tutkimusta varten muodostetuille aineistoille että niiden taustalla oleville alkuperäisille rekisteriaineistoille, lähetetään XML-pohjaisina DDI3.1-tiedostoina.
- kohorttien poiminnassa käytetyt pseudonyymilistat lähetetään raakatekstinä, yksi pseudonyymi rivinä kohden.
- aineistojen eheys tarkastetaan metatietojen perusteella (järkevät arvot).

#### 7.4.9. Pseudonymisointi

Pseudonyymi lasketaan alkuperäisestä tunnisteesta tai muusta tiedosta seuraavalla tavalla:

$$\text{pseudonyymi} = \text{sha256}(\text{kateno}(\text{tunniste}, \text{salalause}))$$

jossa "tunniste" on alkuperäinen, pseudonymisoitava tieto (merkkijonona); "salalause" on projektikohtainen merkkijono, jonka viranomaisen hakee MIDRAS-järjestelmästä; "kateno" on funktio, joka liittää kaksi merkkijonoa peräkkäin; ja "sha256" on funktio, joka laskee merkkijonon SHA-256-tiivisteeseen.

Jokaisella projektilla on oma salalause, jottei pseudonymisointialgoritmia pysty käyttämään kukaan muu kuin aineistoa lähettävät viranomaiset, ja jotta pseudotunnisteet ovat erilaiset joka projektin aineistossa. Kaava pseudonyymien tuottamiseen on periaatteessa yksisuuntainen, eli pseudonyymistä ei pysty laskemaan alkuperäistä tietoa tai tunnistetta, vaikka tuntisikin salalauseeseen; kuitenkin jotkin arvojoukot, kuten mahdollisten henkilötunnusten joukko, ovat niin pieniä, että salalauseen tunteva osapuoli voi muodostaa kaksisuuntaisen muunnoksen muodostamalla kaikki mahdolliset tunnukset ja laskemalla niiden vastaavat

pseudonyymit. Niinpä pseudonymisointiin käytettävä projektikohtainen salalause on tietosuojan kannalta kriittinen.

Kun pseudonymisoituun aineistoon tarvitsee liittää uusia aineistoja esim. aineiston päivittämiseksi uusista tiedoista, poimitaan uudet tiedot vanhan aineiston pseudonyymilistan perusteella.

## 8. Etenemissuunnitelma

### 8.1. Toteutusvaiheet ja aikataulu

MIDRAS-toimintamallin toteutuminen on monivaiheinen prosessi, koska viranomaisten lähtötilanteet ovat erilaiset. MIDRAS-toimintamalli hioutuu ja laajenee ajan myötä, ja tutkimusinfrastruktuuria valmistelevat toimet (palveluiden rakentaminen, valmisaineistojen muodostaminen ja kuvailu) vievät oman aikansa.

#### 8.1.1. Perustamisvaihe (2011-2012)

Perustamisvaiheeseen kuuluvat kaikki ne toimet, jotka pitää tehdä, jotta MIDRAS-järjestelmä saadaan käyttöön. Perustamisvaiheessa MIDRAS toimii ennen kaikkea uutena rekisteritietojen lupahakemus- ja luovutuskanavana. Aineistojen lähetysvaihe aikaisemmassa tietojenluovutusprosessissa on korvattu tietoteknisellä tiedonsiirtoyhteydellä aineiston ylläpitäjän ja MIDRAS-järjestelmän välillä. MIDRAS tarjoaa myös keskitetyn paikan, jossa viranomaiset ja rekisteritutkijat voivat kommunikoida ja johon voi kerryttää tietoa rekisteriaineistoista ja rekisteritutkimuksesta.

Toimet	Vaikutus
MIDRAS-palvelun ensimmäinen tekninen toteutus otetaan tuotantokäyttöön.  Keskeiset viranomaiset liittyvät MIDRAS-järjestelmään ja alkavat tarjota aineistonsa tutkimuskäyttöön MIDRAS-lupaprosessin ja -etätyöpöydän kautta.	Tietoluovutusten tietoturva paranee.  Tutkimusaineistojen korjaus ja päivitys helpottuu.  Kokemus etäkäytöstä alkaa kertyä.
MIDRAS-järjestelmään alkaa kertyä aineistojen kuvailuja ja arkistoituja tutkimusaineistoja.	Tutkijat saavat kuvailut käyttöönsä nopeammin.  Aineistojen kuvailusta kertyy kokemusta.
Käyttölupaa hakevat tutkijat osoitetaan MIDRAS-järjestelmän käyttäjiksi.  MIDRAS-järjestelmään liittyneiden viranomaisten käyttölupaprosessi sähköistetään.	Tutkijat saavat käyttöönsä ajantasaiset, työskentelypaikasta riippumattomat työvälineet  Rekisteritutkimuksen tunnettuus lisääntyy

### 8.1.2. Palveluvaihe (2012-2014)

Palveluvaihetta ovat ne kehitysaskleet, joita MIDRAS-järjestelmässä toteutetaan yhä useampien palveluiden ollessa käytössä. Palveluvaiheessa MIDRAS-järjestelmä sisältää koko ajan laajenevan valikoiman tietoa tarjolla olevista aineistoista, niiden käsittelystä, niillä tehdystä tutkimuksesta ja etäjärjestelmän käytöstä. Viranomaiset ja tutkijat voivat korvata osan omista tietojenkäsittelytarpeistaan MIDRAS-etätyöpöydän tarjoamilla palveluilla. Suuri osa viranomaisista toteuttaa valmisaineistovaraston, johon tiedot tulevat automaattisesti tuotantotietokannoista.

Toimet	Vaikutus
Aineistojen tuottajat muodostavat valmisaineistoja ja automatisoivat aineistoluovutusten prosessia.	Yksittäisen aineistoluovutuksen työmäärä vähenee.
Järjestelmään liittyy uusia aineistojen tuottajia omalähtöisesti tai yhdistelmäaineistojen yhtenä lähteenä.	Saatavilla olevien kuvailtujen aineistojen pohja laajenee.
Aineistojen kuvailuja kerätään ja muunnetaan yhteiseen muotoon.	Tutkimusten suunnittelu, aineistojen löytäminen ja aineistojen ymmärtäminen helpottuu.
Tietojen vaihto MIDRAS-järjestelmän ja muiden aineistokatalogien välillä käynnistyy.	Suomeen saadaan valtakunnallinen infrastruktuuri tutkimuksen tietoaaineistojen hallinnointiin.
Keskitetty etätyöpöytä kehittyy käyttäjäpalautteen perusteella ja alkaa vähentää erillisten tiedonkäsittely-ympäristöjen tarvetta.	Tutkijoiden tietotaito tulee hyödynnetyksi muiden tutkijoiden työssä.
Palveluun aletaan kerätä eri rekistereistä yhdisteltyä saatavuusaineistoa.	Tutkimuskohorttien suunnittelu nopeutuu. Osassa aineistopyynnöistä prosessi yksinkertaistuu.

### 8.1.3. Jatkokehitysvaihe (2014-)

Jatkokehitysvaiheeseen kuuluvat ne toimet, joissa MIDRAS-järjestelmä kehittyy uusiin ja ennalta arvaamattomiin suuntiin. Jatkokehitysvaiheessa MIDRAS-järjestelmä on yleiskäyttöinen tietojen keruu- ja käsittelyalusta, joka soveltuu erityisesti arkaluontoisen tiedon käsittelyyn. MIDRAS-järjestelmän oheispalvelut helpottavat kommunikaatiota tiedon tuottajien ja käyttäjien kesken. MIDRAS-järjestelmässä tehtyä vakiointityötä hyödynnetään muissa palveluissa. Automatisointi lisää jatkuvasti aineistojen ajantasaisuutta ja nopeuttaa lupa- ja aineistonluovutusprosesseja.

Toimet	Vaikutus
MIDRAS-järjestelmän teknisiä ratkaisuja päivitetään.	Ajantasainen, tarpeisiin vastaava palvelu
Viranomaiset rakentavat rajapintoja, joilla voi tehdä tietokyselyitä tutkimuskäyttöön tarkoitettuihin ai-	Aineistotoimitusten kommunikaatioviipeet pienenevät.

neistoihin.	
Tietopyyntöjen tarkistusta (esim. otoskoko) automatisoidaan.	Läpinäkyvä tietojen käyttö vähentää etukäteiskontrollin tarvetta.
Aineistojen kuvailuja linkitetään kasvavassa määrin kansallisiin ja semanttisen webin käsitteistöihin.	Avautuu ennakoimattomia tietojen käyttömahdollisuuksia.
MIDRAS-järjestelmästä saatua rajapintakokemusta hyödynnetään muiden palveluiden kehittämisessä.	Tietojen vaihto lisääntyy ja ratkaisut vakioituvat.
MIDRAS-järjestelmää käytetään sovelluskehitysalustana, kun käsiteltävä tieto on arkaluontoista.	Esimerkiksi sairaanhoidon tietotekniikan kehitysprosessia saadaan kevennetyksi.

Kuva 20 MIDRAS-mallin toteutusehdotus vaiheittain

	2011	2012	2013	2014
<b>Edellytykset</b>	Päätös hallintomallista ja rahoituksesta	Lakimuutokset ja sopimukset		Asenteiden ja tietotekniikan kehitys
<b>MIDRAS-järjestelmän toteutus</b>	Suunnittelu	Ensimmäinen tekninen toteutus tuotantokäyttöön	Kehitys käyttäjien palautteen myötä	
<b>MIDRAS-tukikeskus (hallinto)</b>	Tutkijapalvelu; Suunnittelu		Lakisääteiset tehtävät	
<b>Datakatalogi ja metatiedot</b>	Aineistokatalogin suunnittelu ja katalogien yhteisfoorumi	Järjestelmään alkaa kertyä yhtenäisiä kuvailuja; JHS-suositus	Tiedonvaihto käynnistyy katalogien välillä	Aineistojen kuvailujen linkitys semanttisen webin käsitteistöihin
<b>Sähköinen lupamenettely</b>	Yhteinen lupahakemus	Kehitys käyttäjien palautteen myötä		
<b>Etätyöpöytä ja tukipalvelut</b>		Etätyöpöytä ja tukifoorumi tutkimuskäytössä	Etätyöpöytä tekee tarpeettomaksi osan muista ympäristöistä	Teknisten ratkaisujen päivitys; Sovelluskehitysalusta
<b>Viranomaisten liittyminen MIDRAS-järjestelmään</b>	Keskeiset viranomaiset mukaan		Uusia aineiston tuottajia mukaan	
<b>Aineistot</b>		Aineistoja aletaan toimittaa; Valmisaineistojen muodostaminen	Saatavuusaineistoja eri rekistereistä	
<b>Aineistojen toimittaminen</b>		Aineistoja toimitetaan järjestelmään push-mallin mukaisesti	Aineistoluovutusten automatisoinnin kehitys; Pull-mallin testausta	Tietoluovutuksissa yhteiset rajapinnat

## 8.2. Sidosryhmät ja yhteistyöhankkeet

MIDRAS-selvityshankkeella on useita sidosryhmiä, joiden toiminnalla on MIDRAS-hankkeen kanssa yhteisiä tavoitteita. Nämä sidosryhmät ovat olleet luontevia yhteistyötahoja niin hankkeen aikana kuin tulevaisuudessa. On olennaisen tärkeää, että eri sidosryhmät tekevät yhteistyötä ja osallistuvat MIDRAS-järjestelmän kehittämiseen, koska hyvin monella alalla on käyttöä MIDRAS-järjestelmää tai sen osia muistuttaville järjestelmille. Aktiivisella keskustelulla varmistetaan, ettei päällekkäisiä palveluita synny ja että MIDRAS-järjestelmän kautta voidaan tarjota laaja valikoima aineistoja tutkimuskäyttöön.

### 8.2.1. Tutkimuksen tietoaaineistot –selvityshanke

Opetusministeriö asetti keväällä 2009 poikkihallinnollisen selvityshankkeen kartoittamaan ja koordinoimaan julkisin varoin tuotettujen sähköisten tietoaaineistojen ja tietovarantojen hyödyntämisen tehostamista. Hankkeen tavoitteena oli ohjata kansallista tutkimuksen tietoaaineistoja koskevaa selvitystyötä, muodostaa kokonaiskäsitys Suomen tilanteesta ja laatia kansallinen suunnitelma (tiekartta) tutkimuksen tietoaaineistojen saatavuuden ja säilytyksen kehittämiseksi. Hankkeen ohjausryhmä sai työnsä päätökseen vuoden 2010 lopussa.

Tutkimuksen tietoaaineistot -hankkeessa tuotiin esiin monia tietoaaineistojen käyttöön tällä hetkellä liittyviä haasteita ja tarpeita joita on korostettu myös MIDRAS-hankkeen loppuraportissa. Myös ehdotukset tarvittavista toimenpiteistä ovat linjassa toistensa kanssa. Hankkeen väliraportissa *”Tutkimuksen tietoaaineistot – Olennaisen käsikirja päättäjille”* (CSC – Tieteen tietotekniikan keskus 2010) sekä loppuraportissa *”Tieto käyttöön – Tiekartta tutkimuksen sähköisten tietoaaineistojen hyödyntämiseksi”* (Opetus- ja kulttuuriministeriö, 2011) ehdotetaan monia konkreettisia toimenpiteitä sähköisten tietoaaineistojen, myös rekisteriaineistojen käytön edistämiseksi. Toimenpiteet liittyvät erityisesti kansallisen tietopolitiikan laatimiseen, lainsäädännön tarkistamiseen, toimintamallien kehittämiseen sekä kansallisen tietoinfrastruktuurinkin rakentamiseen. MIDRAS-hankkeeseen olennaisesti liittyviä toimenpide-ehdotuksia ovat esimerkiksi: tietoaaineistojen tutkimuskäytön hinnoitteluperusteiden selventäminen, aineistojen saatavuusperiaatteiden ja niihin liittyvän lainsäädännön tarkistaminen, tilastolain uudistaminen, henkilötietolainsäädännön täsmentäminen, yhteistyöfoorumien perustaminen tietoaaineistojen löydettävyyden takaamiseksi ja yhteentoimivuuden tekniikoiden määrittämiseksi, yhtenäisten käyttöehtojen ja lisenssien kehittäminen, periaatteiden ja käytännösääntöjen luominen julkisille organisaatioille ja tutkimuskentälle tietosuojan alaisten tietoaaineistojen käsittelyyn sekä tutkimuksen tietoinfrastruktuurin kokonaisrakenteen ja yhteisten palvelujen suunnittelu.

Tutkimuksen tietoaaineistot –hankkeen jatko on tällä hetkellä vielä avoinna. Hankkeen loppuraportissa ehdotetaan poikkihallinnollisen koordinaatioryhmän perustamista, joka seuraisi systemaattisesti tutkimuksen

tietoaineistojen saatavuuden ja hyödynnettävyyden edistymistä sekä julkisen sektorin tuottamien ja hallinnoimien tietoaineistojen että tutkimusjärjestelmän tuottamien tietoaineistojen osalta, ja laatisi toimintaehtotuksia asian edelleen edistämiseksi. Mikäli tällainen koordinaatioryhmä perustetaan, on tärkeää, että koordinaatioryhmä huomio kehitystoiminnassa myös rekisteritutkimuksen kehitystarpeet ja MIDRAS-hankkeen yhteydessä kerätyt opit ja kokemukset sekä edistää etäkäyttöpalvelun kehitystä.

Konkreettinen kehitysaskel jota MIDRAS-hankkeen ja Tutkimuksen tietoaineistot –hankkeen tiimoilta lähdetään jo nopeasti edistämään on CSC:n koordinoima yhteistyöfoorumien perustaminen tietoaineistojen löydettävyyden takaamiseksi ja tietoaineistokatalogien yhteentoimivuuden tekniikoiden määrittämiseksi.

### 8.2.2. Biopankkitutkimus

Biopankit ovat ihmisistä otettujen näytteiden (esimerkiksi verinäytteiden) kokoelmia, sekä näihin näytekoelmiin liittyviä tietokantoja, joihin on talletettu näytteille tehtyjen analyysien ja testien tuloksia. Biopankkitutkimus käyttää aineistonaan biopankkien näytteitä, tulostietokantoja tai molempia. Biopankit saattavat sisältää valtavan määrän tietoa, esimerkiksi eri yksilöiden geeniaineistoa. Suomessa on erittäin hyvät ja monipuoliset biopankkiaineistot, ja Suomen voidaankin sanoa olevan maailman huippuluokkaa olemassa olevan aineiston kannalta.

Biopankkitutkimuksella on osittain hyvin samanlaisia tarpeita kuin rekisteritutkimuksella. Useat tutkimukset tarvitsevat infrastruktuuria tietojen yhdistelemiseksi useammista lähteistä. Lisäksi biopankkitutkimus tarvitsee usein taustatiedoksi tavanomaista rekisteriaineistoa, kuten demografista tietoa. Tällä hetkellä biopankkien tutkimuskäytölle luodaan yhteistä infrastruktuuria. Myös MIDRAS-järjestelmä on voi toimia osana biopankki-infrastruktuuria: MIDRAS tarjoaa hyvän yhdistely- ja tutkimusalustan biopankkiaineistoille niiltä osin, kuin tiedot ovat sähköisessä muodossa.

Toisaalta biopankkitutkimuksen tyypilliset tarpeet eroavat jonkin verran rekisteritutkimuksesta. Erityisesti geenitutkijat tarvitsevat suurten tietomääriensä käsittelyyn erikoistuneita ohjelmistoja ja tavallista rekisteritutkijaa enemmän talletustilaa sekä tietojenkäsittelykapasiteettia. Geeniaineistojen analysointiin käytetyt työkalut ovat aktiivisen tutkimuksen kohde, eikä ole helppoa tarjota tutkijoille keskitetysti ohjelmistovalikoimaa, joka vastaisi heidän tarpeisiinsa. Biopankkitutkimuksen tarpeet pystytään kattamaan MIDRAS-ympäristössä tarjoamalla tutkijoille eristetyn MIDRAS-ympäristön sisällä myös omia laskentaympäristöjä, esimerkiksi virtuaalipalvelimia.

Sekä biopankki- että rekisteritutkimuksessa on selkeä tarve julkaista tietoja käytettävissä olevista aineistoista ja koota ne yhteen paikkaan. MIDRAS-järjestelmän metatietokatalogi on yksi mahdollinen julkaisukanava biopankkien metatiedoille.

Biopankkitutkimuksen tukemiseksi ja kehittämiseksi on Suomessa perustettu BioMedInfra-yhteistyöhanke, jossa toimivat BBMRI-yhteistyöryhmä (biopankkiaineistot), EATRIS-hanke (biolääketieteen soveltava tutkimus) ja ELIXIR-hanke (tutkimusta tukeva tietotekniikka). Nämä ovat MIDRAS-järjestelmälle luontevia yhteistyökumppaneita.

### **8.2.3. Yhteiskuntatieteellinen tietoaarkisto**

Yhteiskuntatieteellinen tietoaarkisto on Suomen Akatemian aloitteesta 1999 perustettu tutkimuksen ja opetuksen valtakunnallinen palveluinfrastrukturi. Tietoaarkiston tarkoituksena on arkistoida ja välittää koti- ja ulkomaisia elektronisia tutkimusaineistoja tutkimuksen, opetuksen ja opiskelun käyttöön. Yksikkö toimii Tampereen yliopiston yhteydessä ja sen toimintaa rahoittaa opetus- ja kulttuuriministeriö. Tietoaarkisto on merkittävä yhteiskuntatieteiden tutkimusinfrastrukturi, jolla on myös vankkaa kokemusta yksikköaineistojen metatiedoista. MIDRAS-järjestelmän rajapintoja ja metatietostandardeja tulee kehittää yhdessä tietoaarkiston kanssa.

Tietoaarkiston verkkopalvelussa julkaistaan käytettävissä olevien aineistojen aineistoluettelot, kunkin aineiston metatiedot, muuttujakuvaillut, kyselylomakkeet ja haastattelurungot. Aineistojen käyttö edellyttää käyttöluvuhakemusta ja allekirjoitettua käyttöehtositoumusta. Aineistot toimitetaan vastaanottajalle sähköpostin välityksellä. Tietoaarkistoon arkistoiduista aineistoista on aina poistettu tunnistetiedot. (Yhteiskuntatieteellinen tietoaarkisto, 2011)

Tietoaarkiston tehtävänä on huolehtia aineistojen kuvailusta, pitkäaikais säilytyksestä, konvertoinneista ja jatkokäytön organisoinnista. Lisäksi tietoaarkisto huolehtii teknisestä tietoturvasta ja aineistojen tietosuojasta. Tietoaarkisto on ollut aktiivisesti mukana kehittämässä uutta metadatan kuvailuformaattia DD13.

Yhteiskuntatieteellinen tietoaarkisto kuuluu eurooppalaisten data-arkistojen kattojärjestöön CESSDA:an (Council of European Social Science Data Archives, 2011), jonka hankkeet ovat osa eurooppalaisia infrastruktuurihankkeita. Lisäksi Yhteiskuntatieteellinen tietoaarkisto osallistuu Suomen ainoana osapuolena eurooppalaisten tietoaarkistojen ja tilastoviranomaisten yhteiseen, Euroopan laajuiseen Data Without Boundaries-infrastruktuurin hankkeeseen.

### **8.2.4. TK-online-projekti ja etäkäyttöjärjestelmien yhteiskehittämistä selvittävä työryhmä**

Tilastokeskukseen perustettiin vuonna 2001 tutkimuslaboratorio. Tutkimuslaboratoriossa tutkijat voivat Tilastokeskuksen koneilla ja tiloissa päästä tekemään valvotusti analyysyjä ensisijaisesti yritystietoja sisältävistä aineistoista. Tutkimuslaboratoriotyöskentely on kuitenkin koettu hankalaksi ja alueellisesti epäoikeudenmukaiseksi. Tämän tilannetta on pyritty parantamaan etäkäyttöprojektin avulla.

Tilastokeskuksessa perustettiin keväällä 2008 etäkäyttöprojekti, jonka tavoitteena oli rakentaa ja ottaa käyttöön online-käyttöyhteys Yrityksen rakenteet –yksikössä sijaitsevan tutkimuslaboratorion aineistoihin (Tilastokeskus, 2009). Projektin aikana selvitettiin etäkäyttöjärjestelmän teknisiä ratkaisuja ja ylläpidon mahdollista ulkoistamista, kartoitettiin etäkäyttöjärjestelmän kautta käyttöön tarjottavia aineistoja sekä rakennettiin pilottiympäristö, jota neljän eri tutkimuslaitoksen tutkijat käyttivät. Tilastokeskuksen etäkäyttöprojekti onkin ollut hyvin arvokas tietolähde MIDRAS-hankkeen etäkäyttöjärjestelmän teknisessä suunnittelussa. Myös Tilastokeskuksen työ metatietostandardien suhteen on ollut selvitysohjelmalle arvokasta.

Projektin loputtua syksyllä 2009 Tilastokeskus on hakenut rahoitusta toiminnan jatkamiseksi ja laajentamiseksi. Etäkäyttötoiminta on pysynyt pienessä mittakaavassa käynnissä, mutta toimintaa ei ole mainostettu laajasti. Projektin loppuraportissa ehdotetaan, että etäkäyttöympäristöä tulisi pilottijärjestelmään verrattuna kasvattaa teknisesti kapasiteettiä lisäämällä. Lisäksi Tilastokeskuksen muitakin kuin yritysaineistoja tulisi valmistaa etäkäyttöjärjestelmän kautta käytettäväksi. Projektin loppuraportti löytyy Tilastokeskuksen verkkosivuilta (Tilastokeskus, 2009).

Tilastokeskuksen esittämien suunnitelmien mukaan etäkäyttömahdollisuutta tulisi laajentaa niin, että sen välityksellä olisi saatavilla myös muiden virastojen aineistoja. Selvitettäväksi kuitenkin jää, tulisivatko nämä aineistot tällöin tilastolain alaisiksi, jolloin aineistot pitäisi anonymisoida ennen niiden luovuttamista tutkijoille. Valtiovarainministeriö on asettanut tilastolain muutosta miettivän työryhmän, joka toimii vuoden 2011 loppuun asti (valtiovarainministeriö, 2010b). Tilastolaki (280/2004) ja EU:n tilastoasetus (223/2009) eroavat toisistaan varsinkin niissä kohdissa, jotka koskevat tietojen luovuttamista tutkimuskäyttöön. EU:n tilastoasetuksen mukaan tutkijoille voidaan luovuttaa aineistoja, joissa henkilöt ovat välillisesti tunnistettavissa. Jos tilastolakia muutetaan vastaamaan EU:n tilastoasetusta, tutkimusaineistojen anonymisointi jää työvaiheena pois ja aineistojen luovuttaminen nopeutuu ja halpenee.

Valtiovarainministeriö on asettanut Tilastokeskuksen ja MIDRAS-hankkeen etäkäyttöjärjestelmien yhteistyöryhmän, jonka tarkoituksena on selvittää etäkäyttöohjelmien yhteiskehittämismahdollisuudet (valtiovarainministeriö, 2010c). Lisäksi ryhmän tavoitteena on tehdä ehdotus etäkäyttöjärjestelmien hallintomalliksi sekä järjestelmien toteuttamisen ja ylläpidon resurssoinniksi ja rahoitukseksi. Valtiovarainministeriön asettamien tavoitteiden mukaisesti Suomessa tulisi tarjota hallinnollisten aineistojen sekä tilastorekisterien ja -aineistojen tietoja yhtenäisen etäkäyttöjärjestelmän välityksellä. Tämä järjestelmä olisi osa kansallista tutkimuksen e-infrastruktuuria. Työryhmän työ valmistuu 30.6.2011. MIDRAS-selvityshanke pyrkii tuottamaan työryhmälle tietoa, auttaa työryhmää esittämään kaikkia osapuolia palvelevan mallin etäkäyttöjärjestelmän toteuttamiseksi.

### 8.2.5. RAKETTI-tietovarasto

RAKETTI-XDW on korkeakoulujen ja opetusministeriön yhteinen hanke, jossa rakennetaan tiedolla johtamista tukeva tietovarasto korkeakoulujen ja opetusministeriön tarpeisiin (CSC - Tieteen tietotekniikan keskus, 2011). Tietovarastoon kerätään ajantasaista ja vertailukelpoista tietoa korkeakouluista: opiskelijamääristä, opintosuoritteista, henkilöstöstä, tutkintomääristä, taloustilanteesta ja niin edelleen. Tätä tietoa tarvitaan niin korkeakoulujen sisällä kuin opetus- ja kulttuuriministeriössä.

Tietojen käsittelyyn ja käyttöön on suunniteltu etäkäyttöympäristöä, jossa ministeriön ja korkeakoulujen edustajat voivat käyttää erilaisia raportointityökaluja tuottaakseen tilastotietoa päätöksenteon ja toiminnan suunnittelun tueksi. Etäkäyttöympäristö tarjoaisi tässäkin tapauksessa monipuolisen, helppokäyttöisen ja turvallisen ratkaisun arkaluontoisten tietojen käsittelyyn.

Raketti-tietovaraston suunniteltu etäkäyttöympäristö muistuttaa MIDRAS-etätyöpöytää sekä joiltain aineistotyypeiltään (esim. luottamuksellisia, henkilötietoja sisältäviä aineistoja) että mahdollisesti joiltain työkaluiltaan. Onkin mahdollista, että yksi järjestelmä pystyisi kattamaan molemmat tarpeet: koulutustietojen viranomaiskäytön (RAKETTI) ja yleisesti viranomaisrekisterien tutkimuskäytön (MIDRAS).

### 8.2.6. Julkishallinnon IT-kehitysryhmät

Valtiovarainministeriön ja liikenne- ja viestintäministeriön alaisuudessa toimii työryhmiä, jotka suunnittelevat omalta osaltaan julkishallinnon tietohallinnon periaatteita. Erityisesti valtiovarainministeriön työryhmä "Julkishallinnon perustietovarantojen rajapinnat" (PERA, Valtiovarainministeriö 2010a) tuottaa suosituksia siitä, miten tietoja tarjotaan tietovarannoista. MIDRAS-kehitystyössä pitää ottaa huomioon näiden työryhmien suositukset..

## Lähteet

### Artikkelit, selvitykset ja suositukset

- Borchsenius, Lars (2005). New developments in the Danish system for access to micro data. Invited Paper. Submitted by Statistics Denmark. In: Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, November 9-11 2005.  
[http://www.dst.dk/upload/new\\_access\\_to\\_micro\\_data\\_from\\_statistics\\_denmark\\_001.pdf](http://www.dst.dk/upload/new_access_to_micro_data_from_statistics_denmark_001.pdf)
- CSC – Tieteen tietotekniikan keskus (2010). Tutkimuksen tietoaaineistot – Olennaisen käsikirja päättäjille. Espoo, CSC- Tieteen tietotekniikan keskus. <http://www.csc.fi/csc/julkaisut/oppaat/2010/tutkimuksen-tietoaaineistot>
- Hoeve, Frans (2009). Microdata access in the Netherlands. Statistical Journal of the IAOS, Vol 26 (3), 95-100.
- Johansen, Jon Roy & Litton, Jan-Eric (2005). Security Policies for TwinNET. Draft 4. GenomeEUTwin.  
[http://www.genomeutwin.org/member/docs/Policy\\_document\\_TwinNET.pdf](http://www.genomeutwin.org/member/docs/Policy_document_TwinNET.pdf)
- Kuula, Arja & Borg, Sami (2007). Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari. Valmistelu-  
raportti OECD:n datasuosituksen toimeenpanomahdollisuuksista Suomessa. Yhteiskuntatieteellisen tietoa-  
kiston julkaisuja 6, 2007. [http://www.fsd.uta.fi/julkaisut/julkaisusarja/FSDjs06\\_OECD.pdf](http://www.fsd.uta.fi/julkaisut/julkaisusarja/FSDjs06_OECD.pdf)
- Opetus- ja kulttuuriministeriö (2011). Tieto käyttöön. Tiekartta tutkimuksen sähköisten tietoaaineistojen  
hyödyntämiseksi. Opetus- ja kulttuuriministeriön julkaisuja 2011:4.  
[http://www.minedu.fi/OPM/Julkaisut/2011/Tiekartta\\_tutkimuksen\\_sahkoisten\\_tietoaaineistojen\\_hyodynta\\_miseksi.html](http://www.minedu.fi/OPM/Julkaisut/2011/Tiekartta_tutkimuksen_sahkoisten_tietoaaineistojen_hyodynta_miseksi.html)
- Rekisteritutkimuksen tukikeskus (2010). Rekisterien käyttö väitöskirjoissa 2005–2010. Saatavissa  
[www.rekisteritutkimus.fi](http://www.rekisteritutkimus.fi).
- Reuter, Wolf Heinrich & Museux, Jean Marc (2010). Establishing an Infrastructure for Remote Access to  
Microdata at Eurostat. Teoksessa Domingo-Ferrer J. & Magkos, E. (eds.): PSD 2010, LNCS 6344, 249-257,  
2010. Springer-Verlag Berlin Heidelberg.
- Ritchie, Felix (2009). UK release practices for official microdata. Statistical Journal of the IAOS, Vol 26 (3),  
109-117.
- Tam, Siu-Ming (2009). Official statistics and microdata – access and confidentiality. Statistical Journal of the  
IAOS, Vol 26 (3), 55-56.
- Tam, Siu-Ming, Farley-Larmour, Kim & Gare, Melissa (2009). Supporting research and protecting confiden-  
tiality. ABS microdata access: Current strategies and future directions. Statistical Journal of the IAOS, Vol 26  
(3), 65-74.
- Tilastokeskus (2009). Projektin loppuraportti. Tutkimuslaboratorion etäkäyttöprojekti. Laatija Jouko Verho.  
[http://www.stat.fi/org/tilastokeskus/ya21o\\_etakaytto-loppuraportti.pdf](http://www.stat.fi/org/tilastokeskus/ya21o_etakaytto-loppuraportti.pdf)
- Vetenskapsrådet (2008). Report from the Conference: Global challenges – regional opportunities. How Can  
Research Infrastructure and eScience Support Nordic Competitiveness? 12-13 November 2008.  
<http://www.vr.se/download/18.227c330c123c73dc586800013477/NordicConferenceRInov2008.pdf>

## Lait ja asetukset

Euroopan Parlamentin ja Neuvoston asetus (EY) N:o 223/2009 Euroopan tilastoista. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:fi:PDF>

Laki terveydenhuollon valtakunnallisista henkilörekistereistä (556/1989)

Tilastolaki (280/2004)

## Julkaisemattomat muistiot ja raportit

Edin, P-A., Pedersen, J. & Jonsson, J. (2006). Förslag till samarbete med SCB för förbättrad tillgång till registerdata för forskningsändamål. DISC, Vetenskapsrådet.

Fomkin, Ruslan (2009). Cross-organizational Database Infrastructure for register-based research (CODIR). Edited by Kate Svensson. Phase One Report. DISC, Vetenskapsrådet.

Statistiska Centralbyrån (2003). Slutrapport från Mikrodataprojektet. Microdata On-Line Access System at SCB - MONA.

## Hankkeet ja internet-sivut

Australian Bureau of Statistics (2010). International working group on Microdata Access. <http://www.abs.gov.au/websitedbs/d3310114.nsf/4a256353001af3ed4b2562bb00121564/a924e0190b629c8aca2576b200229b08!OpenDocument>

BioGrid Australia (2010). <http://www.biogrid.org.au/wps/portal>.

Council of European Social Science Data Archives (2011). <http://www.cessda.org/>

CSC – Tieteen tietotekniikan keskus (2011). Raketti. Rakenteellisen kehittämisen tukena tietohallinto. <http://raketti.csc.fi/xdw>

Danmarks Statistik (2008). Om TIMES. <http://www.dst.dk/Vejviser/dokumentation/times/OmTIMES.aspx>

DDI (Data Documentation Initiative) (2009). <http://www.ddialliance.org/>

Dwbproject (2011). Data without Boundaries – DwB. <http://www.dwbproject.org/>

Luxembourg Income Study, (2000). <http://www.lisproject.org/>

Office for National statistics (2008). The virtual microdata laboratory. <http://www.ons.gov.uk/about/who-we-are/our-services/vml>

SDMX (Statistical Data and Metadata eXchange). <http://sdmx.org/>

Statistiska centralbyrån (2011). Microdata för forskare. [http://www.scb.se/Pages/List\\_\\_\\_\\_257147.aspx](http://www.scb.se/Pages/List____257147.aspx)

Statistiska centralbyrån (2008). Dokumentation med MetaPlus. <https://www.h2.scb.se/metadatas/begrepp.aspx>

UK Data Archive (2011). <http://www.data-archive.ac.uk/>

Valtiovarainministeriö, 2010a. Julkishallinnon perustietovarantojen rajapinnat –työryhmä.  
[http://www.hare.vn.fi/mHankePerusSelaus.asp?h\\_ild=15360](http://www.hare.vn.fi/mHankePerusSelaus.asp?h_ild=15360)

Valtiovarainministeriö, 2010b. Tilastolain uudistamista valmisteleva työryhmä. Asettamispäätös 30.8.2010.  
[http://www.hare.vn.fi/mHankePerusSelaus.asp?h\\_ild=16575](http://www.hare.vn.fi/mHankePerusSelaus.asp?h_ild=16575)

Valtiovarainministeriö, 2010c. Tilastokeskuksen ja Midras-projektin etäkäyttöjärjestelmien yhteiskehittämistä valmisteleva työryhmä. Asettamispäätös 31.5.2010.

Yhteiskuntatieteellinen tietoarkisto (2011). <http://www.fsd.uta.fi/>

## Liitteet

### Liite 1. Sanasto

#### **Aggregointi**

Yhden aineiston yksikkötietojen yhdistäminen toisiinsa siten, että tuloksena oleva aineisto sisältää summattua tietoa (frekvensseinä tai tunnuslukuina) lähdeaineiston tiedoista

#### **Aineistokuvaus**

Metatieto, jossa selvitetään, miten aineisto on hankittu, mistä se koostuu jne.

#### **Anonymisointi**

Aineiston käsitteleminen siten, että aineiston yksiköiden sekä suora että välillinen tunnistaminen ovat mahdottomia

#### **Autentikointi, käyttäjätunnistus**

Jotain kommunikointireittiä (esim. tiettyä tietoliikenneyhteyttä) käyttävän ihmisen varmistaminen tietyn käyttäjätunnuksen haltijaksi, ja sitä seuraava käyttäjätunnuksen merkitseminen tätä kautta käytettyjen resurssien ja palveluiden käyttäjäksi

#### **Federaatio**

Usean organisaation tai tietojärjestelmän yhteenliittymä

#### **Federoitu tietokanta (virtuaalinen tietokanta)**

Tietokanta, jonka tietosisällöt säilytetään useammassa taustatietokannoissa

#### **Henkilörekisteri**

Käyttötarkoituksensa vuoksi yhteenkuuluvista merkinnöistä muodostuva henkilötietoja sisältävä tietojoukko, jota käsitellään osin tai kokonaan automaattisen tietojenkäsittelyn avulla taikka joka on järjestetty kortistoksi, luetteloksi tai muulla näihin verrattavalla tavalla siten, että tiettyä henkilöä koskevat tiedot voidaan löytää helposti ja kohtuuttomitta kustannuksitta

- henkilötiedolla tarkoitetaan kaikenlaisia luonnollista henkilöä taikka hänen ominaisuuksiaan tai elinolosuhteitaan kuvaavia merkintöjä, jotka voidaan tunnistaa häntä tai hänen perhettään tai hänen kanssaan yhteisessä taloudessa eläviä koskeviksi.

#### **Kokonaisaineisto**

Rekisterin kaikista yksilöistä/yksiköistä muodostuva aineisto

#### **Käyttäjätunnus (tekninen termi)**

Henkilökohtainen tunniste, johon liittyvillä tunnistamistiedoilla (esimerkiksi salasana ja kännykän numero vahvaa tunnistamista varten) käyttäjä pystyy todistamaan tietojärjestelmälle, kuka hän on.

- Käyttäjätunnus myönnetään käyttöluvan saaneelle henkilölle (käyttölupasopimuksen tekijäosapuolelle) käyttöluvassa määritellyn aineiston käyttämistä varten

#### **Käyttö lupa (hallinnollinen termi)**

Anomuksen käsittelyn tuloksena syntyvä sopimus, jonka perusteella käyttäjällä (MIDRAS-järjestelmässä tutkijalla) on oikeus käyttää jotain aineistoa tai sen osaa

#### **Käyttöoikeus (tekninen termi)**

Käyttäjätunnukseen liitetty tieto siitä, mitä resursseja kyseistä tunnusta käyttävä käyttäjä saa käyttää; Käyttöluvan tekninen toteutus

#### **Käyttökopio, tekninen kopio**

Sellainen kopio tiedosta, jota ei pidetä erikseen yllä (voidaan tuottaa aina automaattisesti uudelleen).

## **Luokitus**

Lista tiettyjä arvoja saavan muuttujan arvojen merkityksistä

## **Matriisi**

Aineisto tai aineiston osa, joka sisältää kaikista aineiston käsittelemistä yksilöistä samat tiedot (muuttujat)

## **Metatieto**

Tiedon kontekstia, sisältöä ja rakennetta sekä niiden hallintaa ja käsittelyä koko elinkaaren ajan kuvaavaa tietoa. Tätä tietoa voidaan käyttää mm. aineiston hakuun, paikallistamiseen ja tunnistamiseen. Metatiedot ovat olennaisia aineistojen löytämisen, luetteloinnin ja käytön kannalta. Metatiedoista saatetaan käyttää myös termiä metadata. Metatiedot sisältävät sekä aineiston kuvailutietoja että teknisiä, järjestelmän metatietoja.

## **Mikrodata, yksilöaineisto**

Yksikköaineisto, jossa havaintoyksikkönä on henkilö

## **Muuttuja**

Määrämuotoisen aineiston yksi sarake eli tieto, joka annetaan aineiston jokaiselle riville eli havainnolle

## **Otos**

Yksikköaineistosta tuotettu versio, jossa on vain osa yksiköistä (henkilöistä, yrityksistä tai muista rekisteriyksiköistä) mukana. Käytetään, jos esimerkiksi rekisteriin liittyvät lait tai käytännöt estävät kokonaisaineiston tutkimuskäytön.

## **Primäärirekisteri**

Hallinnollinen alkuperäinen tietokanta, jossa rekisteridata sijaitsee

## **Pseudonymisointi**

Tunnisteiden korvaaminen yksiselitteisillä tunnisteilla, joista alkuperäistä tunnistetta ei voi päätellä

## **Pseudotunniste, koodi**

Pseudonymisoinnissa (de-identifioinnissa) alkuperäisen tunnisteiden korvaava uusi tunniste.

## **Pull-malli**

Tiedonsiirtokäytäntö, jossa tiedon lähettäjä tarjoaa tiedonsiirtorajapinnan ja vastaanottaja avaa tiedonsiirtoyhteyden

## **Push-malli**

Tiedonsiirtokäytäntö, jossa tiedon vastaanottaja tarjoaa tiedonsiirtorajapinnan ja lähettäjä avaa tiedonsiirtoyhteyden

## **Rajapinta**

Määrittely siitä, miten jonkin järjestelmän, palvelun tai ohjelmiston kanssa voidaan kommunikoida; Joskus myös tämän määrittelyn toteuttava ohjelmisto/komponentti

## **Rekisteri**

Yksikötasoista aineistoa sisältävä tietokanta, jota tuottaa ja ylläpitää rekisterinpitäjä, ja joka on talletettu rekisterinpitäjän tietojärjestelmään tai tietojärjestelmiin

- Rekisteri sisältää yksilöivän tunnisteiden, esimerkiksi henkilötunnus tai Y-tunnus, jolla rekisterin yksi yksikkö voidaan tunnistaa.
- Yleensä rekisteri on jatkuvasti päivittyvä hallinnollinen tietokanta, ei kertaalleen kerätty tai tilastotarkoitukseen kerätty hetkellinen tietokanta.

## **Rekisteriaineisto**

Rekisteriin pohjautuva eli rekisteritiedoista muodostettu aineisto

## **Rekisterinpitäjä**

Yksi tai useampi henkilö, yhteisö, laitos tai säätiö, jonka käyttöä varten henkilörekisteri perustetaan ja jolla on oikeus määrätä henkilörekisterin käytöstä tai jonka tehtäväksi rekisterinpito on lailla säädetty. Lainsäädännöllisesti myös henkilötietoja sisältävän tutkimusaineiston saanut tutkija on rekisterinpitäjä.

## **Rekisteriseloste**

Henkilötietolain 10 § edellyttämä julkinen metatieto henkilörekisteristä

## **Saatavuusaineisto**

Yksikötason aineisto (mikrodata), joka sisältää tietoa lähinnä siitä, mitä tietoja on saatavissa mistäkin yksiköstä. Voidaan käyttää esim. tutkimusta tukevien biopankkiaineistojen löytämiseen.

## **Tietojärjestelmä**

Tietokanta tai data warehouse -ratkaisu, jossa rekisteritiedot on talletettu rekisterinpitäjän hallinnoimana.

## **Tietokanta (Database)**

Tietotekniikassa käytetty termi tietojen varastolle. Se on määrämuotoinen kokoelma tietoja, joilla on yhteys toisiinsa. Käytännössä tietokanta yleensä viittaa nykyisin ns. relaatiotietokantaan ja johonkin tietokantaohjelmistoon, jolla tietokannan sisältöä käsitellään.

## **Tietokantafederaatio**

Useista eri lähteistä saatavien tietosisältöjen tarjoaminen yhtenä tietokantana.

- aidossa tietokantafederaatiossa keskeistä on se, että dataa ei → kahdenneta, vaan se haetaan järjestelmistä tarpeen mukaan

## **Tietovaranto**

Hallinnolliselta tai sisällölliseltä kannalta yhtenäinen kokonaisuus hyödynnettävää tietoa

## **Tietovarasto (Data Warehouse)**

Laaja kokonaisuus (repository), johon on talletettu sähköisesti organisaation tai tietyn aihepiirin tietoa. Tietovarasto on tietokantaa laajempi käsite. Tyypillisesti tietovarastototeutus voi sisältää useita tietokantoja sekä erilaisia tiedon käsittelyyn tarkoitettuja kerroksia (ohjelmistoja).

## **Tilastoaineisto**

Aineisto, jonka käyttötarkoitus on tilastotarkoitus, ei tutkimustarkoitus.

## **Tilastorekisteri**

Primäärirekistereistä tuotettu tilastotarkoitukseen tarkoitettu rekisteri. TK: (tilasto)rekisteri on jatkuvasti päivittyvä aineisto, joka sijaitsee tiedontuottajalla.

## **Tutkimusaineisto**

Tiettyä tutkimusta varten tuotettu (kerätty, muokattu) aineisto, jonka perusteella tutkimus tehdään

## **Tutkimusjoukko, kohortti**

Niiden yksilöiden/yksiköiden joukko, joita tietyssä tutkimuksessa tutkitaan.

## **Vahva tunnistaminen**

Kuten lakitekstissä: autentikointi, jossa käyttäjän varmistaminen käyttäjätunnuksen haltijaksi perustuu ainakin kahteen näistä kolmesta: (1) käyttäjän tietämään asiaan, (2) käyttäjän hallinnoimaan esineeseen, (3) käyttäjän yksilöivään ominaisuuteen

## **Valmisaineisto**

Aineisto, josta organisaatio tuottaa tutkijoille luovutettavia aineistoja.

## **Valmisaineistovarasto**

Tietojärjestelmä, joka sisältää kaikki valmisaineistot

### **Välillinen tunniste**

Ominaisuusryhmä, joka riittää yksiselitteisesti määrittämään aineiston yhden yksilön, esim. osoite+ikä lähes aina tai vaikkapa sukupuoli+hiusten väri+silmien väri+pituus+paino+korvien muoto+jokaisen sormen pituus

### **Yhdistely, linkkaus**

Kahden aineiston yhdistäminen käyttäen avaimena aineistoista löytyviä tunnisteita (esim. henkilötunnus)

## **Liite 2. Vaihtoehtoiset mallit ja ratkaisut**

Tässä liitteessä käsitellään ne erilaiset lähestymistavat ja mallit rekisteriaineistojen etäkäytöstä, jotka otettiin huomioon selvitystyötä tehtäessä ja joiden soveltuvuutta ja kannattavuutta tutkijoiden, viranomaisten ja muiden sidosryhmien kannalta hankkeessa on selvitetty. Vaihtoehtojen selvitystyö on toiminut taustamateriaalina MIDRAS-ehdotusta koottaessa.

### **Selvityksessä tarkastellut vaihtoehtoiset aineistontuotantomallit**

Yleisellä tasolla voidaan nähdä kolme toimintatapaa:

#### *5. Yksittäiset tutkimusaineistot*

Aineistot muodostetaan käsityönä tuotantorekistereistä aina pyyntöjen tullessa.

#### *6. Valmisaineistovarasto*

Tutkimuspyyntöjä varten pidetään yllä automaattisesti päivittyviä, tutkimuskäyttöön soveltuvia kokonaisaineistoja, joista tietopyynnöt toteutetaan tarpeen mukaan rajattuina.

#### *7. Tuotantorekisterien suora käyttö*

Tutkijoiden käyttöön annetaan suoraan tuotantorekisterin tiedot tarpeen mukaan rajattuina.

Suosittu toimintamalli vastaa vaihtoehtoa 2. Tarkoituksena on, että tietojen luovuttaminen tutkimusaineistoiksi olisi automatisoitavissa mahdollisimman pitkälle, kunhan käyttöluvahakemus on hyväksytty. Suositamme valmisaineistojen muodostamista, koska ainakin yksinkertaiset tietopyynnöt on mahdollista toteuttaa tällaisista aineistoista automaattisesti sähköisen lupahakemuksen perusteella ja toisaalta erillinen tutkimusrekisteri on sisällöltään vakaampi ja tietoturvan kannalta helpompi hallita kuin varsinainen tuotanto- eli primäärirekisteri. Suositamme myös, että viranomaisen tekee tutkimusaineistojen muodostus- ja muokkaustyöstä vain sen osan, joka väistämättä kuuluu viranomaiselle, kuten tietojen kerääminen useammista tuotantorekistereistä ja aineiston karkea rajaus käyttöluvahakemuksen perusteella. Tutkijan tehtäväksi sen sijaan jätetään muun muassa johdettujen muuttujien muodostus (eli tietojen päättelemine rekisteriaineistosta) sekä mielenkiintoisten tapausten ja verrokkien poimiminen. Aineiston oikean tulkinnan pitää perustua kunnollisiin metatietoihin, mukaan lukien kuvailu siitä, miten aineisto on muodostunut.

## Selvityksessä tarkastellut lupahakemuksen toimintamallivaihtoehdot

Selvityksessä on tarkasteltu kolmea vaihtoehtomallia:

1. *Sähköinen yhteinen lupahakemus – luvat annetaan hajautetusti*

Sähköinen lupahakemusjärjestelmä jakaa kopion lupahakemuksesta jokaiselle viranomaiselle, jonka tietoja lupahakemuksessa pyydetään. Viranomainen käsittelee hakemusta, ja ilmoittaa MIDRAS-hallinnolle käyttöluvan saaneista.

2. *Sähköinen yhteinen lupahakemus – luvat annetaan keskitetysti*

Säädetään erityinen lakipykälä, joka antaa yhdelle taholle valtuudet myöntää tutkimuslupia muiden omistamiin aineistoihin (nykyisen tilastolain tapaan).

3. *Manuaalinen hajautettu malli*

Jokainen viranomainen toteuttaa oman versionsa lupahakemuksesta, nykytilanteen mukainen malli.

Vaikka käyttöluvan myöntämisprosessi nopeutuisi selkeästi, jos yhdellä taholla olisi valtuudet päättää kaikkien viranomaisten rekisterien käyttöluvista, ei Suomessa olla vielä valmiita tällaiseen ratkaisuun. Viranomaiset ovat pitäneet perusolettamuksena, että heillä säilyy vastuu päättää aineistojen luovutuksista.

## Selvityksessä tarkastellut vaihtoehtoiset pseudonymisointimallit

Selvitystyössä tutkittiin, missä vaiheessa pseudonymisointi olisi syytä tehdä ja kenen vastuulla se on. Selvitetävänä oli kolme vaihtoehtoa:

1. *Keskitetty malli*

Kaikki aineistot tarjotaan tunnisteellisina MIDRAS-järjestelmään, jossa ne pseudonymisoidaan ennen tutkijalle luovuttamista.

2. *Hajautettu malli, satunnainen salalause*

Jokaiselle projektille muodostetaan satunnainen salalause tutkimusprojektia perustettaessa. Viranomainen hakee salalauseen MIDRAS-järjestelmästä (tai viranomaiset sopivat salalauseen keskenään). Aineistot tarjotaan MIDRAS-järjestelmään tällä salalauseella pseudonymisoituina.

3. *Hajautettu malli, johdettu salalause*

Sovitetaan viranomaisten kesken kaavasta, jolla uusien tutkimusprojektien salalauseet muodostetaan. Aineistot tarjotaan MIDRAS-järjestelmään projektille muodostetulla salalauseella pseudonymisoituina.

Taulukko 12 Eri pseudonymisointivaihtoehtojen vertailu

ominaisuus	keskitetty malli	hajautettu, satunnainen	hajautettu, johdettu
aineistot siirtyvät pseudonymisoituina verkossa		X	X
ei tarvetta varjella jotain tahojen "yhteistä salaisuutta"	X	X	
ei tarvetta pseudonymisointiavainten lähettämiseen tahojen välillä	X		X
vain yhden tahon tarvitsee toteuttaa pseudonymisointi	X		
käytetyn avaimen ei tarvitse olla MIDRAS-järjestelmässä		X	X
pöimintaan käytettävät kohortit voidaan välittää pseudonyymeina		X	X

Hajautettu toimintamalli satunnaisin projektikohtaisin salalausein katsottiin tietoturvan kannalta parhaaksi toteutustavaksi. Lisäksi pilotoinnin ohessa todettiin, että pseudonymisointialgoritmin käyttö aineistoa lähettäville viranomaisille ei muodosta olennaista lisätaakkaa aineistotoimituksille. Johdettujen salalauseiden toimintatapa todettiin erityisen ongelmalliseksi siksi, että se vaatii jonkinlaisen rajanvedon siitä, mille tahoille yhteinen pseudonymisointiavainten muodostustapa voidaan paljastaa, eli ketkä voivat olla aineiston tuottajia MIDRAS-toimintamallissa.

### Selvityksessä tarkastellut vaihtoehtoiset pseudonymisointitekniikat

MIDRAS-järjestelmässä tutkija ei saa koskaan käyttöönsä suoria tunnisteita sisältäviä aineistoja, vaan suorat tunnisteet (erityisesti henkilötunnukset) korvataan tutkimuskohtaisilla surrogaateilla, pseudonyymeilla. Pseudonyymien tuottamiseksi punnittiin kahta erilaista menetelmää:

#### 1. Perinteinen pseudonymisointiavain

Kuvaus (taulu), joka yhdistää kunkin tutkimusprojektissa mukana olevan tunnisteiden vastaa vaan pseudonyymiin. Pseudonyymit ovat tyypillisesti numeroita yhdestä ylöspäin.

#### 2. Algoritminen pseudonymisointi salalauseella

katakin tunnistetta vastaava pseudonyymi tuotetaan tunnisteesta tunnetulla algoritmilla, joka määritellään MIDRAS-järjestelmän laajuisesti. Jokaisella projektilla on kuitenkin oma salalauseensa, jota käytetään osana pseudonyymien tuottamista siten, että projektien pseudonyymit ovat erilaisia eikä pseudonyymeja pysty muodostamaan tuntematta salalauseita. Salalause on vain projektiin aineistoja toimittavien tahojen tiedossa.

Taulukko 13 Pseudonymisointivaihtoehtojen vertailu

ominaisuus	perinteinen ratkaisu	algoritminen ratkaisu
pseudonyymit palautettavissa helposti alkuperäisiksi tunnisteiksi	X	
henkilötunnusta vaihtaneet yksilöt voi yhtenäistää pseudonymisoinnissa	X	

kohortin henkilötunnuslistaa ei tarvitse tietää etukäteen		X
pystyy pseudonymisoimaan myös muuta tunnistedataa		X
henkilötunnuslistoja ei tarvitse lähettää organisaatioiden välillä		X
organisaatiot pystyvät pseudonymisoimaan aineistot toisistaan riippumatta		X

Algoritminen pseudonymisointi on osoittautunut pilotoitaessa kevyemmäksi toimintamalliksi, ja tietoturvakriittisissä tilanteissa salalauseen sopiminen aineiston tuottajien kesken on helpompaa kuin perinteisen pseudonymisointiavaimen jakaminen, mikä edellyttää käytännössä aina tiedostojen lähettämistä tavalla tai toisella. On huomattava, että koska pseudonymisointialgoritmillä voi pseudonymisoida mitä vain (ei pelkään esim. henkilötunnuksia tai y-tunnuksia), tunnisteiden eheystarkistus on tehtävä ennen pseudonymisointia.

### Selvityksessä tarkastellut aineistotoimitusten vaihtoehtoiset mallit

Selvityshankkeessa tutkittiin viittä eri mallia siitä, miten viranomaisen tarjoaa aineistot MIDRAS-järjestelmään, eli miten MIDRAS-järjestelmän ja aineiston tuottajien välinen tietofederaatio on ratkaistu. Eri mallit eivät ole keskenään poissulkevia, vaan MIDRAS-järjestelmä voi tukea useampia yhtäaikaan. Vaihtoehtoina olivat:

#### 1. Kertaluontoinen tiedonsiirto

Rekisterinpitäjä muodostaa aineiston kerran, ja aineisto sekä sen käyttöoikeudet lisätään MIDRAS-järjestelmään ihmistyönä kertaluontoisesti.

#### 2. Yleiskäyttöinen kertaluontoinen tiedonsiirto

Aineisto lisätään MIDRAS-järjestelmään ja päivitetään siellä ihmistyönä, mutta sen käyttötarkoitus on avoin ja rekisterinpitäjä voi lisätä siihen uusia käyttöoikeuksia ja poistaa vanhoja ilman, että näistä tarvitsee erikseen sopia tai tehdä ihmistyötä.

#### 3. Automaattinen tiedonsiirto, push-malli

MIDRAS-järjestelmä tarjoaa rekisterinpitäjälle keinon lisätä, päivittää ja poistaa aineistoja ilman MIDRAS-ylläpidon ihmistyötä sekä antaa käyttöoikeuksia näihin.

#### 4. Automaattinen tiedonhaku järjestelmätason tunnistuksella, pull-malli

Rekisterinpitäjä tarjoaa MIDRAS-järjestelmälle keinon hakea aineistoja (myös päivityksiä, uusia aineistoja) ilman rekisterinpitäjän ihmistyötä. Käyttöoikeuksia hallinnoidaan MIDRAS-järjestelmässä.

#### 5. Automaattinen tiedonhaku käyttäjätason tunnistuksella, pull-malli

Rekisterinpitäjä tarjoaa MIDRAS-järjestelmälle keinon hakea aineistoja (myös päivityksiä, uusia aineistoja) ilman rekisterinpitäjän ihmistyötä. Käyttöoikeuksia hallinnoidaan MIDRAS-järjestelmässä ja rekisterinpitäjän järjestelmässä.

Taulukko 14 Aineistotoimitusvaihtoehtojen vertailu

ominaisuus	kertaluontoinen tiedonsiirto	yleiskäyttöinen tiedonsiirto	automaattinen tiedonsiirto	tunnukseton tiedonhaku	tunnuksellinen tiedonhaku
aineistojen lisäys, päivitys ja eheystarkistus automatisoitavissa			X	X	X
aineistoa käyttävien tutkimusprojektien lisäys automatisoitavissa		X	X	X	X
muistuttaa perinteistä toimintamallia, tietoluovutus viranomaiselta lähtöisin	X	X	X		
mahdollista hakea aineistot MIDRAS-järjestelmään tarpeen mukaan				X	X
mahdollista käyttää aineistopalvelua muussa tiedon jakelussa			(X)	X	X
aineiston tuottaja voi seurata aineistoon tehtyjä kyselyjä				X	X
aineiston tuottaja myös osaltaan tarkistaa aineistoon tehdyt kyselyt					X
valtiohallinnon tietotekniikka-arkkitehtuurin suositusten mukainen			(X)	X	X

MIDRAS-suositus on toteuttaa MIDRAS-järjestelmään vaihtoehdot 3 ja 5, joista viranomaisille suositetaan käytettäväksi vaihtoehtoa 5. Pilotoinnin yhteydessä todettiin, että tietoteknisesti automatisoidut tiedonsiirtotavat eivät muodosta olennaista lisävaivaa aineistotoimituksissa vaan pikemminkin vähentävät vaivaa ja auttavat virheiden korjaamisessa. Automatisoidussa tiedonsiirrossa on kuitenkin omat virhemahdollisuutensa, joiden vuoksi tietojen on hyvä tarkistaa aineiston eheys MIDRAS-järjestelmässä metatietojen perusteella. Viranomaisten täytyy perustaa vaihtoehdossa 5 aineistoja tarjoava palvelu, jonka vuoksi vaihtoehto 3 sopii välivaiheen ratkaisuksi.

## Selvityksessä tarkastellut metatietohallinnon vaihtoehtoiset mallit

Selvitystyön ohessa on kartoitettu, millaisia metatietoja aineistojen tutkimuskäyttö edellyttää; tämä metatiedon vaatimusmäärittely on esitetty liitteessä 6. Lisäksi tarvitsee sopia, missä muodossa metatiedot esitetään. Selvityksessä on tutkittu seuraavat vaihtoehdot:

### 1. Vapaamuotoinen metatietosäilö

Metatiedot talletetaan MIDRAS-järjestelmään siinä muodossa, kuin ne sattuvat olemaan. Metatietoja etsitään esim. vapaiden tekstihakujen perusteella.

### 2. Vakiintunut metatietoformaatti

Käytetään olemassaolevaa standardia tiedostomuotoa metatietojen keräämiseen, esim. DD12, DD13, SDMX, Tilastokeskuksen PXML, RIF-CS tai CSMD.

### 3. MIDRAS-järjestelmän metatietoformaatti

Määritetään uusi tiedostomuoto tai välitystapa metatiedoille, jota käytetään MIDRAS-järjestelmässä; tiedostomuoto voi pohjautua esimerkiksi raakatekstiin, XML-merkintäkieleen tai RDF-merkintäkieleen.

Vapaamuotoiset metatiedot tuottavat vähiten vaivaa metatietojen tuottajille, mutta niiden jatkokäytettävyyden on huono, organisointi vaikeaa ja vaara metatietosäilön tietojen väärinymmärryksistä suuri. Vakiintunutta metatietoformaattia kannattaa käyttää mieluummin kuin itse määriteltyä, jos vain jokin vakiintuneista metatietoformaateista pystyy vastaamaan MIDRAS-järjestelmän vaatimuksiin. DD13 on tällainen standardi ja vahvin kandidaatti yhteiseksi kansainväliseksi mikrodatan kuvailustandardiksi. Muut vakiintuneet metatietoformaattit eivät vastaa MIDRAS-järjestelmän vaatimusmäärittelyä. Valitettavasti DD13:n käsitteilyyn ja tuottamiseen ei ole vielä olemassa käyttäjälähtöistä dokumentaatiota ja työskentelyvälineitä, joten MIDRAS-järjestelmän on tarjottava aineistojen tuottajille tällaiset välineet.

## Selvityksessä tarkastellut vahvan tunnistuksen vaihtoehdot

Selvityksessä on otettu huomioon seuraavat tunnistautumistekniikat:

### 1. Sirukorttintunnistautuminen

käyttäjä tunnistetaan sirukortille (esim. väestörekisterikeskuksen organisaatiokortille) asennetulla varmenteella, jonka voi avata asettamalla kortin kortinlukijaan ja syöttämällä sirukortin PIN-numeron.

### 2. SMS passcode

käyttäjä tunnistetaan ensin käyttäjätunnus-salasanaparilla. Tämän jälkeen käyttäjän matkapuhelimeen lähetetään kertakäyttöinen salasana, jolla käyttäjä pääsee palveluun.

### 3. Kertakäyttösalausanaalista

käyttäjille lähetetään etukäteen lista kertakäyttösalasanoja, joista aina annetaan yksi kirjautumisen yhteydessä käyttäjätunnus-salasanaparin lisäksi.

#### 4. Mobiilivarmenne

käyttäjä tunnustetaan matkapuhelimen SIM-kortille asennetulla varmenteella, jonka voi avata antamalla puhelimelle PIN-koodin jonkin palvelun pyytäessä varmistusta.

Taulukko 15 Vahvan tunnistuksen vaihtoehtojen vertailu

ominaisuus	sirukortti	SMS code	pass-	salasanalista	<u>mobiilivarmenne</u>
käyttäjällä yleensä valmiiksi tarvitut välineet	(X)	X			
helppokäyttöinen tekniikka	(X)	X		X	X
kypsä, vakiintunut tekniikka	(X)	X		X	
käyttäjien identiteetin varmistaa luotettu kolmas taho	X				(X)

SMS passcode valittiin ennen kaikkea siksi, että sen edellyttämä laiteinfrastruktuuri on valmiiksi olemassa eikä edellytä tunnistautumisvälineiden lähettämistä käyttäjille säännöllisesti. On mahdollista, että tulevaisuudessa mobiilivarmenne tai sirukortti ja niiden käyttövälineet yleistyvät niin paljon, että näistä tulee kannattavampia tunnistautumisvaihtoehtoja.

### Selvityksessä tarkastellut tutkimusympäristövaihtoehdot

Tutkijoiden kannalta MIDRAS-järjestelmän ensisijainen palvelu on varsinainen rekisteritietojen etäkäyttöympäristö - aineistojen käyttöä tukevan metatietopalvelun ohella. Etäkäyttöympäristön tärkein tavoite on vastata tarpeisiin, joita tutkijoilla on aineistojen analysoinnissa ja käsittelyssä tutkimuksiaan varten. Selvityshankkeessa arvioidut etäkäyttötekniikat olivat:

#### 1. Etätyöpöytä

Ne etäkäyttötekniikat, joissa aineistoja työstetään vuorovaikutteisesti palvelimella. Palvelimen työpöytä tarjolla olevine ohjelmistoinen näkyy käyttäjälle, ikään kuin työpöytä ja ohjelmat olisivat käyttäjän paikallisella koneella.

#### 2. WWW-käyttöliittymä

Tavallisen WWW-selaimen avulla käytetty yhteys, jossa käyttäjä analysoi tutkimusaineistojaan erilaisten WWW-palveluun rakennettujen toimintojen avulla. Osa näistä toiminnoista voi olla jonkinlaisia käyttöliittymiä analyysiohjelmiin.

#### 3. Eräajokäyttöliittymä

Esimerkiksi sähköpostilla tai hyvin suoraviivaisella WWW-palvelulla toteutettu järjestelmä, jossa käyttäjä lähettää haluamansa analyysit jollain kielellä kirjoitettuna ohjelmana palvelimelle. Palvelimella ohjelma suoritetaan ja sen tulokset lähetetään takaisin käyttäjälle.

**Taulukko 16 Tutkimusympäristövaihtoehtojen vertailu**

ominaisuus	etätyöpöytä	www-käyttöliittymä	eräajokäyttöliittymä
monipuolinen työkaluvalikoima	X		
voi käyttää omia ohjelmia	X		
välitön palaute tutkijalle	X	X	
tuttu työskentely-ympäristö	X		
vähäinen kehitystyö	X		X
hyvin rajattu käyttäjätuki		X	X
tietoturvan vaatimat käyttörajoitukset helppoja asettaa		X	X
laaja asiakasohjelmakanta	(X)	X	X

Etätyöpöytäratkaisu on monipuolisin ja vastaa useimpiin tutkijatarpeisiin. Muut ratkaisut ovat kuviteltavissa, mutta vaativat käyttäjiltä enemmän opettelua eivätkä ole yhtä helposti laajennettavissa kattamaan uusia tutkimusmenetelmiä. Vaihtoehdot eivät sulje toisiaan pois, mutta jos etätyöpöytäratkaisu otetaan käyttöön, muiden vaihtoehtojen tarpeellisuus vaikuttaa kyseenalaiselta. Niinpä MIDRAS-toimintamallissa suositetaan pelkkää etätyöpöytää.

### Selvityksessä tarkastellut etätyöpöytätekniikat

Etätyöpöytätekniikka tarkoittaa niitä verkkoprotokollia ja ohjelmistoja, joilla käyttäjät muodostavat etätyöpöytäyhteyden MIDRAS-järjestelmän palvelimelle. Selvityksessä kartoitettiin alla olevassa taulukossa esiteltäviä protokollia ja niiden toteutuksia.

**Taulukko 17 Tarkastellut etätyöpöytätekniikat**

etäyhteysprotokolla	asiakasohjelmat	huomautuksia
RDP	Windows Remote Desktop Connection, rdesktop	Tuki yhteyksien vahvalle salakirjoitukselle lisätty RDP-versiossa 6.0
ICA	Citrix Receiver / ICAclient	palvelinpuolen ohjelmisto kalliimpi kuin RDP-protokollalla
RFB	TightVNC	Ei sisäänrakennettua tukea yhteyksien salakirjoitukselle

Käyttäjän kannalta erot etätyöpöytätekniikoiden välillä ovat vähäiset. RDP valittiin etäyhteyksien protokollaksi, koska asiakasohjelma löytyy valmiiksi Windows-koneista ja on helposti asennettavissa Mac- ja Linux-

koneisiin. ICA-protokolla on vakiintuneempi ja tarjoaa jonkin verran monipuolisemmat mahdollisuudet kuin RDP-protokolla, mutta näille lisäominaisuuksille ei ole MIDRAS-järjestelmässä erityistä käyttöä.

## Selvityksessä tarkastellut etätyöpöydän aineistomuotovaihtoehdot

Selvityksessä otettiin huomioon seuraavat muodot, joissa aineisto voi olla tarjolla etätyöpöydällä:

### 1. Tietokanta

Tutkimusaineistot ovat saatavilla erillisessä SQL-tietokannassa, johon otetaan yhteyttä etätyöpöydältä. Tutkimusaineisto haetaan tietokannasta sitä käyttäviin ohjelmiin SQL-kyselyllä esim. ODBC-rajapinnan kautta.

### 2. Tilasto-ohjelman talletusmuoto

Tutkimusaineistot talletetaan tutkijan etätyöpöydälle tilasto-ohjelman talletusmuodossa, esim. SAS-tiedostoina.

### 3. Vakioitu talletusmuoto

Tutkimusaineistot talletetaan tutkijan etätyöpöydälle hyvin tunnetussa, vakiossa talletusmuodossa, esim. csv-tiedostoina.

Taulukko 18 Aineistomuotovaihtoehtojen vertailu

ominaisuus	tietokanta	tilasto-ohjelmatiedosto	vakiotiedosto
rajapintakomponentit ilmaisia		(X)	X
helppokäyttöisyys	(X)	X	X
johdetut näkymät aineistoon	X		
metatiedot aineiston mukana	(X)	X	
aineistoa voi käyttää useammalta etätyöpöydältä	X		
helppo automaatio MIDRAS-järjestelmässä	X		X
mahdollista hakea tiedot tarpeen mukaan	X		

Tietokanta on aineistojen luovutustapana ilman muuta monipuolisin. Vaikka SQL-tietokantojen käyttö on monimutkaisempaa kuin raakojen aineistotiedostojen, monet tilasto-ohjelmat sisältävät aputoimintoja tietojen hakemiseksi SQL-tietokannoista. Lisäksi erillisellä tietokantapalvelimella oleva aineisto on käytettävissä tarvittaessa useammalta etätyöpöydältä, erilaisista tietojenkäsittely-ympäristöistä ja eri tavoin johdettuina näkyminä. SQL-tietokanta tarjoaa tutkijoille sellaisenaankin tärkeän analyysityökalun.

Järjestelmän jatkokehityksen kannalta on myös tärkeää, että tutkija hakee tiedot jonkinlaisen rajapinnan kautta, jos tiedot haetaan MIDRAS-järjestelmään vasta tutkijan niitä tarvitessa. SQL-kyselykieli on ainoa riittävän vakiintunut ratkaisu sen määrittämiseen, mitä tietoja tutkija haluaa.

## Selvityksessä tarkastellut rajapintavaihtoehdot

Rajapintavaihtoehdot tarkoittavat sitä, millä tietoteknisillä välineillä ja protokollilla yhteydet muodostetaan ja tiedot siirretään, eli minkälainen on viranomaisen ja MIDRAS-järjestelmän välinen tiedonvälitysrajapinta. Selvitystyössä on otettu huomioon seuraavat tiedonvälitystekniikat ja kartoitettu niiden soveltuvuutta MIDRAS-käyttöön:

### 1. Määrämuotoinen sähköposti

Tiedot lähetetään sähköpostissa esimerkiksi tietynmuotoisena liitteenä. Salaus hoidetaan erillisellä, sähköpostiviestin sisällön salaavalla tekniikalla, esim. PGP, S/MIME.

### 2. Tiedostonsiirtoprotokollat

Tiedot haetaan tai lähetetään jollain protokollalla, joka on suunniteltu tiedostojen siirtämiseen järjestelmien välillä. Mahdollisia salattuja protokollia ovat esim. SFTP ja SCP.

### 3. www-rajapinta (web service)

Tiedot haetaan tai lähetetään www-palvelurajapinnan kautta erikseen esim. WSDL-määrittelykielellä määrättyllä tavalla. Välittävänä salattuna protokollana toimii HTTPS.

### 4. Tietokantarajapinta

Tiedot haetaan viranomaisilta tietokantayhteyden kautta. Tietokantayhteydet salataan erikseen esim. SSL-tunneloimalla.

Taulukko 19 Rajapintavaihtoehtojen vertailu

ominaisuus	sähköposti	tiedostonsiirto	www-rajapinta	tietokantayhteys
mahdollinen push-mallissa	X	X	X	
mahdollinen pull-mallissa		(X)	X	X
laajasti tunnettu, vakiintunut ratkaisu	X	X	(X)	
valtionhallinnon arkkitehtuurin mukainen		X	X	
viranomaisilla käytettävissä valmiita käyttäjähallintokomponentteja			X	X
ei sidottu tiettyihin taustajärjestelmiin	X	X	X	
ei suurta tarvetta rajapinnan lisämäärittelyille		X		X

Www-rajapinnat tarjoavat vakioratkaisuista monipuolisimmat mahdollisuudet eivätkä sido tuoteriippuvaiisiin ratkaisuihin, joten niiden käyttö on valittu MIDRAS-suositukseksi. Rajapinnan määrittely edellyttää vielä www-rajapinnan tarkennusta. Määrittämällä rajapinta sopivasti viranomaisen on mahdollista toteuttaa yksinkertainen aineistopalvelu (pull-mallissa) tavanomaisilla HTTP(S)-palvelinohjelmistoilla.

Rajapintojen suunnittelussa olisi voitu myös ottaa huomioon, millainen rajapinta mahdollistaisi tietokantafederaation eri organisaatioiden välillä. Federoidut eli virtuaaliset tietokannat (esim. InfoSphere Federation

Server) ovat kuitenkin tällä hetkellä vielä kypsyttömiä ratkaisuja, joiden rajapintaratkaisujen pysyvyyteen ja yleiskäyttöisyyteen ei voi luottaa. On oletettavaa, että tietokantafederaation yleistyessä myös tietojen välittäminen tietokantojen välillä www-rajapintojen kautta tulee yhdeksi vakioratkaisuksi. Tiedonhaun integrointi tutkijan tietokantakäyttöliittymän kanssa on yksi järjestelmän jatkokehitysalue.

### Selvityksessä tarkastellut vaihtoehtoiset tiedostomuodot

Selvityksessä tarkasteltuja luovutettavien aineistojen ja näiden aineistojen metatietojen toimitusmuotoja ovat:

1. Tietojen lähetyksen vakiintuneessa XML-muodossa (esim. SDMX)
2. Tietojen lähetyksen MIDRAS-järjestelmää varten kehitetyssä XML-muodossa
3. Aineiston lähetyksen CSV-muodossa ja rakenteellisten metatietojen lähetyksen vakiintuneessa XML-muodossa
4. Tietojen lähetyksen jonkin tilasto-ohjelman talletusmuodossa (esim. SPSS portable / SAS)

Käytetty tiedostomuoto ei vaikuta olennaisesti järjestelmän kehitys- ja ylläpitovaivaan. Keskeisimmäksi kriteeriksi muodostui tiedostomuodon vakiintuneisuus ja laaja tuki ilmaisille työkaluille tiedostomuodon käsittelemiseksi, josta syystä CSV-muoto valittiin MIDRAS-suositukseksi. CSV ei kuljeta aineistosta edes rakenteellisia metatietoja mukanaan, joten aineiston mukana täytyy lisäksi lähettää jonkinlainen aineistokuvailu. Tämän kuvailun muodostaminen on todennäköisesti viranomaisille vähemmän työlästä, jos se on samassa muodossa kuin muutkin aineistojen metatiedot MIDRAS-järjestelmässä.

### Liite 3. Keskeiset tutkimuksessa käytettävät rekisterit ja rekistereihin perustuvat aineistot

Seuraavaan taulukkoon on kerätty yhteiskunta- ja terveystutkimuksen kannalta tärkeimpiä valtakunnallisia kokonaisrekistereitä ja rekistereihin perustuvia tietokantoja. Luetellut rekisterit ovat pääsääntöisesti henkilörekistereitä ja ne sisältävät henkilöitä yksilöivän tunnistetiedon. Henkilötunnuksen avulla näiden rekistereiden tiedot voidaan yhdistää toisiinsa. Tarkemmat kuvaukset rekistereistä ja niiden tietosisällöstä löytyvät rekisteriselosteista, joista useimmat on julkaistu asianomaisen rekisteripitäjän internet-sivuilla. Rekisteriselosteiden linkit löytyvät ReTkin verkkosivuilta ([www.rekisteritutkimus.fi](http://www.rekisteritutkimus.fi)).

Taulukko 20 Keskeiset tutkimukseen käytetyt rekisterit

Rekisterit rekisterinpitäjän mukaan	Rekisterin sisältö
<b>Eläketurvakeskus (ETK)</b>	
Ansaintarekisteri (työsuhderekisteri)	Työsuhdetietoja työeläkevakuutuksen piiriin kuuluvista yksityisen sektorin työntekijöistä ja yrittäjistä
Eläkerekisteri	Tietoja yksityisen ja julkisen sektorin työeläkelaitosten antamista eläkepäätöksistä
<b>Fimea</b>	
Lääkkeiden haittavaikutusrekisteri	Tietoja lääkkeiden ja hammaslääkkeiden ilmoittamista epäillyistä lääkkeiden haittavaikutuksista
<b>Kansaneläkelaitos (KELA)</b>	
Etuusrekisteri (tallennettu lukuisiin erillisiin eri etuuksia koskeviin osarekistereihin)	Tietoja liittyen etuuksiin esimerkiksi lapsenhoidosta, sairaudesta, toimeentulosta, koulutuksesta ja työttömyydestä
<b>Maa- ja metsätalousministeriö</b>	
Maatilarrekisteri	Maatilojen keskeisimmät tiedot
Maatilayritysrekisteri	Maatilayritysten perustietoja
<b>Oikeusrekisterikeskus</b>	
Konkurssi- ja yrityssaneerausrekisteri	Konkurssi- ja yrityssaneerausasiatietoja tuomioistuini- ja viranomais-toimintaa sekä velkojen ja sivullisten oikeuksien turvaamista varten
Rikosrekisteri	Valtakunnallinen lakisäänteinen keskusrekisteri rikosoikeudellisten seuraamusten määräämistä ja täytäntöönpanoa varten sekä henkilön luotettavuuden tai henkilökohtaisen soveltuvuuden selvittämistä ja arviointia varten
<b>Opetushallitus</b>	
Ammatillisten oppilaitosten ja lukioiden yhteishakurekisteri	Tietoja toisen asteen ammatilliseen koulutukseen ja lukiokoulutukseen opiskelijaksi hakeutuneista ja opiskelijavalinnoista
Ammattikorkeakoulujen yhteishakurekisteri ja ammattikorkeakoulujen hakija- ja opiskelijavalintojen suorittamisesta ja opiskelupaikan vastaanottamisesta	Tietoja ammattikorkeakoulujen opiskelijaksi hakeutumisesta, opiskelijavalintojen suorittamisesta ja opiskelupaikan vastaanottamisesta

lupaikkarekisteri	
Yliopistojen haku ja opiskeluoikeusrekisteri	Yliopisto- opiskelijaksi hakeutumisen ja opiskelijavalintojen suorittamisen yhteydessä kerättyjä tietoja
<b>Säteilyturvakeskus (STUK)</b>	
Annosrekisteri	Tietoja säteilyaltistuksen seurannassa olevien työntekijöiden saamista säteilyannoksista
<b>Terveyden ja hyvinvoinnin laitos (THL)</b>	
Hoitoilmoitusrekisteri (Hilmo)	Diagnoosi- ja hoitotoimenpidetietoja sairaaloiden osastoilla hoidossa olleista potilaista
Lastensuojelurekisteri	Kodin ulkopuolelle sijoitettujen lasten ja nuorten tietoja
Sosiaalihuollon hoitoilmoitusrekisteri	Tietoja sosiaalihuollon laitoshoidosta ja asumispalvelusta, kotihoitosta, laitosten toiminnasta ja niiden asiakkaista
Syntyneiden lasten rekisteri	Tieto jokaisesta Suomessa syntyneestä lapsesta ja muita taustatietoja
Syöpärekisteri (ylläpitään Suomen Syöpäyhdistys)	Tiedot syöpätautien esiintyvyydestä, ennalta ehkäisystä, hoidosta ja lääkinällisestä kuntoutuksesta sekä palveluiden käytöstä
Tartuntatautirekisteri	Lääkärien ilmoittamien tartuntatautilain määräämien sairauksien ja mikrobiologisten laboratoriodien tiettyjen mikrobilöydösten tiedot
Toimeentulotukirekisteri	Varsinaisen ja ehkäisevän toimeentulotuen saajien tiedot ja kuntouttavan työtoiminnan tukea saaneiden tiedot
<b>Tilastokeskus</b>	
FLEED-aineisto (Finnish Longitudinal Employer-Employee Data)	Työnantajien ja työntekijöiden tiedoista yhdistetty aineisto (henkilötiedot, toimipaikkatiedot ja yritysten tilinpäätöstiedot)
Kuolemansyytilaston aineisto (kuolemansyyrekisteri)	Tietoja kuolleista henkilöistä (esim. kuolinaika ja peruskuolemansyy)
Palapeli-aineisto (parisuhteet, lapset, perheet, elinolot -aineisto)	Aineisto kaikkien Suomen väestöön vuodesta 1970 lähtien kuuluneiden parisuhde-, lapsi-, perhe- ja elinolotiedoista
Tulonjakotilaston aineisto	Tietoja henkilöiden ja kotitalouksien vuositulojen jakautumisesta sekä käytettävissä olevien tulojen määrästä ja muodostumisesta
Tutkintorekisteri	Henkilön suorittamien tutkintojen koulutusaste ja -alatiedot, myös suorittamisajankohta ja oppilaitos
Työssäkäyntitilaston aineisto	Vakituksella vuoden lopussa maassa asuneiden henkilöiden tiedot esimerkiksi pääasiallisesta toiminnasta, ammattiasemasta, työpaikan toimialasta ja suoritetuista tutkinnoista
Väestölaskentojen pitkäikäisaineisto	Tietoja henkilöistä, perheistä, asutokunnista, asunnoista ja rakennuksista (yhdistelty eri vuosien väestölaskenta-aineistoista)
Yritysrekisteri	Yritysten osoitteet, toimialat, henkilöstön ja liikevaihdon suuruusluokat, aloittamisaika sekä tuoja/viejä-tieto
<b>Työ- ja elinkeinoministeriö</b>	

Työhallinnon asiakaspalvelurekisteri (URA)	Työnhakijaksi työvoimatoimistoon ilmoittautuneiden tiedot
<b>Työterveyslaitos (TTL)</b>	
Työperäisten sairauksien rekisteri (ammattitautirekisteri)	Vakuutuslaitoksille ilmoitettujen uusien ammattitautien tiedot ja lääkärin työsuojelun piiriviranomaiselle ilmoittamien uusien ammattitautien ja työstä johtuvien muiden sairaalloisten tilojen tiedot
<b>Verohallitus</b>	
Verotietokanta	Tietoja verotettavien tuloista, vähennyksistä, varallisuudesta ja veloista
<b>Väestörekisterikeskus (VRK)</b>	
Väestötietojärjestelmä	Perustiedot Suomen kansalaisista ja Suomessa vakinaisesti asuvista ulkomaalaisista sekä tietoja rakennuksista, rakennushankkeista, huoneistoista ja kiinteistöistä

## Liite 4. Henkilötietojen luovutusta säätelevät lait

Alla olevassa taulukossa on esitelty keskeisimmät lait, joissa säädetään henkilörekisterissä olevien tietojen luovutuksesta. Henkilötietolaki on aina otettava huomioon henkilöaineistojen eli henkilötietoja sisältävien aineistojen käsittelyssä. Henkilötiedoiksi katsotaan henkilön suorien tunnistetietojen (kuten henkilötunnus, nimi ja osoite) lisäksi kaikenlaiset luonnollista henkilöä taikka hänen ominaisuuksiaan tai elinolosuhteitaan kuvaavat merkinnät, joista hänet tai hänen kanssaan yhteisessä taloudessa elävät ihmiset voidaan tunnistaa.

Taulukko 21 Henkilötietojen luovutusta säätelevät lait

Laki, johon luovutus perustuu	Rekisterit, joiden luovutusta laki säätelee	Mihin tarkoitukseen tietoja voidaan luovuttaa	Erityistä
Henkilötietolaki (523/1999)	Kaikki henkilötietoja sisältävät rekisterit	Historiallista tai tieteellistä tutkimusta sekä tilastointia varten	Saa luovuttaa ilman rekisteröidyn suostumusta --- tietojen suuren määrän, tietojen iän tai muun sellaisen syyn vuoksi; henkilötietojen käsittely perustuu asianmukaiseen tutkimussuunnitelmaan ja tutkimuksella on vastuullinen johtaja; toimitettava niin että tiettyä henkilöä koskevat tiedot eivät paljastu ulkopuolisille.
Julkisuuslaki (621/1999)	Kaikki julkishallinnon tiedot, esimerkiksi Kelan rekisteritietojen luovutukset perustuvat kokonaan julkisuuslakiin.	Tieteellisiin tutkimuksiin, tilastointiin, viranomaisen suunnittelu- ja selvitystyötä varten	Tunnisteet voi luovuttaa, jos henkilötietolain (523/1999) edellytykset täyttyvät.
Tilastolaki (280/2004)	Tilastokeskuksen aineistot (paitsi kuolinsyytietojen luovutukset), THL:n sosiaalirekisteritietojen luovutukset	Tieteellisiin tutkimuksiin ja tilastollisiin selvityksiin	Henkilötietolain mukaisia henkilötietoja ei luovuteta ja siksi tiedot anonymisoidaan niin, ettei välillinenkään tunnistaminen ole mahdollista; ainoastaan tiedot iästä, sukupuolesta, ammatista ja koulutuksesta luovutetaan tunnistetietoineen, jos henkilötietolain (523/1999) edellytykset täyttyvät.
Laki terveydenhuollon valtakunnallisista henkilörekistereistä (556/1989)	THL:n rekisterit, esimerkiksi hoitoilmoitusrekisteri ja syntymärekisteri	Tieteellisiin tutkimuksiin	Tietosuojavaltuutetulla on oikeus tulla kuulluksi luovutus päätöksissä.

Laki kuolemansyyntä selvittämistä (459/1973)	Kuolintodistusarkiston tiedot	Tieteelliseen tutkimukseen, tilastointiin, viranomaisten suunnittelu- ja selvitystyötä varten	
Laki verotustietojen julkisuudesta ja salassapidosta (1346/1999)	Verohallinnon tiedot	Tieteellistä tutkimusta, tilastointia ja viranomaisen suunnittelu- tai selvitystyötä varten	Osa Verohallituksen rekisteritiedoista on julkisia.
Laki väestötietojärjestelmästä ja Väestötietokeskuksen varmenpalveluista (661/2009) ja Valtioneuvoston asetus väestötietojärjestelmästä (128/2010)	Väestötietojärjestelmän tiedot	Yhteiskunnassa yleisesti hyväksytyihin käyttötarkoituksiin	Tiedot ovat salassa pidettäviä, lukuun ottamatta yksittäin luovutettavia osoitetietoja.
Euroopan Parlamentin ja Neuvoston asetus 223/2009/EY Euroopan tilastoista	EU:n tilasto-ohjelman mukaisia tilastoaineistoja	Tutkijoille, jotka suorittavat tilastollisia analyysejä tieteellisiä tarkoituksia varten	Ero tilastolakiin: Viranomaiset voivat toimivaltansa puitteissa myöntää käyttöoikeuden sellaisiin salassa pidettäviin tietoihin, joiden perusteella tilastoyksiköt on mahdollista tunnistaa ainoastaan välillisesti.

## Liite 5. Organisaation ja MIDRAS-ylläpidon välinen esimerkkisopimus

### Sopimus rekisteriaineistojen tutkimuskäyttöpalvelusta etätyöpöydällä

Tämän sopimuksen osapuolet ovat palvelun tuottaja X (MIDRAS-palvelun ylläpitäjä) ja palvelun tilaaja X (organisaatio).

#### Palvelu

Palvelun tuottaja tuottaa rekisteriaineistojen tutkimuskäyttöpalvelun (myöhempänä palvelu), jolla palvelun tilaaja voi luovuttaa aineistojaan tutkimuskäyttöön. Palvelu koostuu eristetystä tietojenkäsittely-ympäristöstä, jossa tutkijat voivat käsitellä aineistoja lataamatta niitä omille koneilleen. Palvelu toteutetaan Windows-etätyöpöytänä, jolta tietojen siirto ulos on estetty. Aineiston käyttöön liittyvät samat rajoitukset kuin tavallisissa aineistoluovutuksissa, eli aineisto annetaan käyttöön palvelun tilaajan myöntämän tutkimusluvan perusteella. Tutkija kirjautuu sisään vahvalla tunnistuksella, jolla varmistetaan käyttäjän olevan henkilö, joka on saanut tutkimusluvan aineistoon.

Aineistot siirretään palvelun tuottaja järjestelmään aineistojen vastaanottopalvelun kautta, jota palvelun tuottaja pitää yllä osana palvelua. Palvelun tuottaja antaa palvelun tilaajalle tunnukset, joilla voi siirtää aineistoja palveluun. Palvelun tilaaja määrittää aineistosiiirron yhteydessä, minkä tutkimusprojektin aineisto on kyseessä, ja palvelun tuottaja tarjoaa sen vain kyseisen tutkimusprojektin tutkijoiden käyttöön palvelun tilaajan myöntämän käyttöluvan mukaisesti. Palvelun tilaaja myöntää käyttöluvat ja tunnistaa tutkijat omien käytäntöjensä mukaan.

#### Osapuolten vastuut

Palvelun tuottaja huolehtii, että aineistoja käsitellään palvelussa palvelun tilaajan erikseen antamien ohjeiden, aineistoja koskevien lakien ja asetusten, hyvän tietohallintotavan ja VAHTI-suositusten 4/2001, 2/2003, 10/2006, 3/2008, 3/2009 ja 2/2010 mukaisesti. Palvelun tuottaja huolehtii, että aineistoja pystyvät palvelussa käyttämään vain henkilöt, joilla on tutkimusluvan perusteella niihin käyttöoikeus. Palvelun tuottaja huolehtii palvelun teknisestä ylläpidosta, tietoturvan tarkkailusta, seurannasta ja raportoinnista palvelun tilaajalle. Palvelun tilaajalla on oikeus auditoida palvelu.

Palvelun tuottaja huolehtii, että aineistojen käsittely on rajattu teknistä ylläpitoa varten tarvitulle henkilökunnalle. Palvelun tuottaja huolehtii, että kaikki aineistoja palvelun tuottajan organisaatiossa käsittelevät henkilöt antavat kirjallisen sitoumuksen tietojen salassapidosta. Palvelun tuottaja ilmoittaa aineistoja käsittelevien henkilöiden nimet ja henkilövaihdokset viipymättä palvelun tilaajalle. Palvelun tuottaja huolehtii,

että aineistoja käsittelevä henkilökunta on perehdytetty aineistoa koskeviin lakeihin ja ohjeisiin. Palvelun tuottaja huolehtii, että aineistot ja kaikki mahdollisesti niistä syntyneet kopiot tai niiden tietoja sisältävät tiedostot poistetaan palvelusta tutkimusluvan päätyttyä.

Palvelun tilaaja huolehtii, että järjestelmään toimitetut aineistot sisältävät ne ja vain ne tiedot, joihin tutkijoiden saama tutkimuslupa myöntää oikeuden. Palvelun tilaaja määrittää palvelun tuottajalle, keillä tutkijoilla on oikeus käyttää aineistoja. Palvelun tilaaja tiedottaa aineistojen käsittelyohjeisiin tulleet muutokset palvelun tuottajalle.

Molemmat osapuolet nimeävät organisaatiostaan yhteyshenkilön, joka huolehtii sopimuksen asianmukaisesta noudattamisesta organisaatiossa.

### **Vastuuhenkilöt**

Palvelun tuottajan organisaatiossa palvelusta vastaa (*henkilön nimi*) ja käyttäjätuesta (*henkilön nimi*). Palvelun tilaajan organisaatiossa palvelun käytöstä vastaa (*henkilön nimi*).

### **Voimassaoloaika**

Tämä sopimus on voimassa allekirjoituksesta lähtien ja jatkuu (*päivämäärä*) saakka. Sopimus voidaan irtisanoa puolin ja toisin välittömästi. Sopimuksen päättyessä palvelun tuottajan on viipymättä poistettava palvelun tilaajan aineistot palvelusta.

### **Allekirjoitukset**

## Liite 6. Metatiedon vaatimusmäärittely

Tässä liitteessä määritellään, mitä tietoja aineistojen kuvailuista tarvitaan tutkimuskäytön kannalta ja kuinka keskeistä mikin tieto on. Määrittelyssä oletetaan, että aineisto on yksi taulu (eli matriisi), joka sisältää kullekin riville samat muuttujat. Jotkin metatiedot liittyvät aineistoon kokonaisuutena, jotkin aineiston yhteen muuttujaan, ja lisäksi muuttujan luokittelussa käytetyillä koodeilla on omat metatietonsa. Jos aineisto sisältää useampia datatauluja (eli relaatioita, matriiseja, sidottuja muuttujajoukkoja), pitää jokainen kuvailla omana aineistonaan.

### Määrittelytaulukon selitteet

Alla olevassa taulukossa on selitetty, mitä määrittelytaulukoiden sarakkeet tarkoittavat.

Taulukko 22 metatietotaulukoiden termien merkitys

Määrittelytaulukoiden selitteet:	
nro	lyhyt nimi, jolla tähän metatietoon voi viitata keskustelussa yms.
nimi	metatiedon sisällön lyhyt kuvaus
määrä	kuinka monta kertaa kyseinen metatieto voidaan antaa kuvattua yksikköä (metatieto, aineisto/taulu, muuttuja, koodi) kohden; "*" tarkoittaa "kuinka monta vain", ja jos tietoa voi olla 0 kpl, tieto on käytännössä valinnainen
kuvaus	määrittely sille, millaista metatietoa tarkoitetaan eli mitä tämän metatietokentän sisällön tulisi olla, myös viittaukset standardeihin, joissa on vastaavia metatietokenttiä
esimerkki	esimerkki siitä, millaista sisältöä kyseisellä metatietokentällä voisi olla jonkin kuvitteellisen aineiston metatiedoissa
tärkeys	luokittelu siitä, kuinka olennainen metatieto on metatiedon tavoitteiden kannalta. MUST: aineiston/metatiedon hyödynnettävyys vaarantuu, jos tietoa ei ole annettu; SHOULD: tieto on hyödyllinen suurimmalle osalle metatiedon käyttäjistä; MAY: tieto on hyödyllinen, mutta ei täytä kumpaakaan edellisistä kriteereistä. Metatiedot, jotka pystyy tuottamaan automaattisesti aineistosta tai aineiston tallennusjärjestelmästä, on merkitty tärkeys-sarakkeeseen merkinnällä <b>(a)</b> .

### Metatiedon tiedot

Jokaisella aineistolla on yhdet metatiedot, ja metatiedon tiedot kuvaavat tätä metatietoa.

Taulukko 23 metatietoa kuvailevat tiedot

Nro	Nimi	Määrä	Kuvaus	Esimerkki	Tärkeys
MT1	julkisuus	1	metatietojen levittämisen rajoitukset	julkinen	SHOULD

MT2	standardi	1	viittaus määritelmään, jonka mukaisesti metatiedot on annettu, myös versio	MIDRAS-metatieto v. 3	MAY (a)
MT3	muutospäivä	1	milloin metatiedot on viimeksi päivitetty	1.9.2007	MUST (a)
MT4	muutostieto	0-*	miten metatiedot ovat muuttuneet ajan mittaan	versio 2: lisätty aineiston lähteitä	SHOULD
MT5	versio	1	versiotunnus, joka tunnistaa tämän aineiston nämä metatiedot yksiselitteisesti	2	MAY
MT6	lähde	1-*	metatiedon tuottaja(t), myös tuottamismenetelmä	MIDRAS	MAY
MT7	kieli	1-*	metatietojen määrittelyyn käytetty kieli	suomi	MAY

## Aineiston tiedot

Aineiston tiedot kuvaavat aineistotaulua kokonaisuutena. Monet aineistolle määritellyt metatiedot voi antaa myös yksityiskohtaisemmalla tasolla (esimerkiksi sekä aineistosta että aineiston sisältämästä yksittäisestä muuttujasta voi olla käyttöoikeustietoa).

Taulukko 24 aineistoa kokonaisuutena kuvailevat tiedot

Nro	Metatieto	Määrä	Selitys	Esimerkki	Tärkeys
A1	nimi	1	lyhyt kuvaus, mitä aineisto on	lapsimäärän pitkittäistilasto	MUST
A2	ylläpito-organisaatio	1-*	aineiston hallinnoija		MUST
A3	yhteystieto	1-*	keneltä saa lisätietoja aineistosta	esim. vastuuhenkilön puhelinnumero	MUST
A4	havaintoyksikkö	1	millaista todellisuuden osaa (yksilöä) jokainen rivi aineistosta vastaa	henkilö	MUST
A5	kattavuus	1	kriteeri, jolla aineistoon kuuluvat yksilöt on valittu	vakituisesti Suomessa asuvat ihmiset	MUST
A6	otos	1	millainen osa kohteesta sisältyy aineistoon	satunnaisotos 1/3	MUST
A7	koko	1	aineiston rivien määrä (eli kuinka monta yksilöä aineisto käsittelee)	2 miljoonaa	SHOULD (a)
A8	muuttujamäärä	1	kuinka monta tietoa (kenttää, muuttujaa) aineisto sisältää yhdestä yksilöstä	15	SHOULD (a)
A9	kuvaus	1	pitkä kuvaus, mitä aineisto on		SHOULD
A10	aikajakso	1-*	millä aikavälillä aineisto on kerätty: alkuperäisen aineiston tuottamispäivämäärien väli	vuodet 1998-2006	MUST
A11	päivitysväli	1	kuinka usein aineistoa päivitetään (aineiston tuottaja määrittää itse,	ei päivitetä	SHOULD

			milloin kyse on uudesta aineistosta ja milloin päivitetystä vanhasta aineistosta; päivityksessä A23 versio muuttuu)		
A12	muutostieto	0-*	useassa erässä tai jatkuvasti kerätyn aineiston keräystavassa, merkinnässä tai tulkinnessa tapahtuneet muutokset liitettynä ajankohtaan, jolloin käytäntö on vaihtunut; yksittäisen muuttujan sisällön tuottamisessa tapahtuneet muutokset merkitään muuttujan metatietoihin (M5)	v. 2003 alkaen otettu huomioon myös vanhempien kuolintieto	SHOULD
A13	käsittely	0-*	mitä aineistolle on tehty/tehdään ennen luovuttamista; jokainen aineistoa jotenkin muuntava lisää A13-elementin, mitä on tehty	osoitemuuttuja poistettu	SHOULD
A14	käyttöoikeudet	1-*	kriteerit, joilla aineistoon annetaan käyttöluvia; tarkemmat, muuttujakohtaiset oikeudet merkitään muuttujan metatietoihin (M12)	annetaan tutkimuskäyttöön ilman henkilötunnuksia	SHOULD
A15	tilausprosessi	1	selitys, miten aineisto on saatavissa, myös käsittelyaika ja asiaanliittyvät maksut	esim. linkki tilauslomakkeeseen ja sen lähetysosoite, n. 3kk, ilmainen	SHOULD
A16	luovutusmuoto	1-*	media ja tiedostomuoto, jossa aineisto luovutetaan käyttöön, myös merkistö	gzip-pakattu SPSS portable -tiedostomuoto (.por.gz) CD-levyllä	SHOULD
A17	sarja	0-1	tiettyyn sarjaan kuuluvasta aineistosta sarjan nimi	suomen demografiset perustilastot	MAY
A18	asiasanat	1-*	aineiston aihepiirit asiasanaston (YSA?) mukaan	lapset, elämäntapa, perheet, sisarukset, syntyvyys, väestönkehitys, väestörakenne	MAY
A19	lähde	1-*	aineiston tuottaja(t), myös kyselyn tekijät, tietojen kerääjät, sekä muut tilastot / rekisterit, joista tietoja on tuotu; yhteenveto muuttujien (M9) tiedoista	synnytyssairaalat, kuolinsyyrekisteri	SHOULD
A20	keruumenetelmä	1	miten tiedot on tuotettu; yhteenveto muuttujien (M9) tiedoista	syntymä- ja kuolinilmoitusten perusteella	MAY
A21	luontipäivä	1	milloin aineiston tuottaja on tuottanut aineiston ensimmäisen kerran	1.5.2007	MAY
A22	muutospäivä	1	milloin aineiston tuottaja on viimeksi päivittänyt aineiston	1.5.2007	MUST (a)
A23	versio	0-1	useaan kertaan päivitetystä aineistosta version nimi, jonka perusteella saman aineistoversion voi löytää myöhemmin	1	SHOULD
A24	käyttötarkoitus	1-*	lähdeaineistoissa olevien tietojen ensisijainen käyttötarkoitus (esim.	tutkimus	MAY

			hallinnollisen rekisterin tietojen käyttötarkoitus); tarkemmat muuttujakohtaiset tiedot voi antaa muuttujien tiedoissa (M18)		
A25	viitteet	0-*	muut aineistot, oheismateriaalit ja tärkeät lähteet, jotka liittyvät aineistoon	esim. linkki ohjekirjaan tai viite tieteelliseen julkaisuun, jossa aineistoa on käytetty	MAY
A26	julkisuus	1	julkisuusaste ja sisältääkö aineisto henkilötietoja	käyttö rajoitettu, sisältää henkilötietoja	SHOULD
A27	kieli	0-*	aineiston vapaamuotoisessa tekstissä käytetty kieli	suomi	MAY (a)
A28	sijainti	1-*	aineiston fyysisen tallennusvälineen sijaintipaikka	Helsinki	MAY
A29	tunnus	1	aineiston tuottajan antama aineiston tuottajalle yksiselitteinen, uniikki tunniste: aineistotietokannan OID, URI, ISBN tms.	355917-20909-1	SHOULD (a)
A30	viittausmalli	1	malli siitä, miten aineistoon viitataan tutkimuksen lähdeluettelossa	Hallinnointiorganisaatio: kansalaisten perustiedot, 2007	SHOULD

## Muuttujien tiedot

Aineistotaulussa on sarakkeita, joista jokainen edustaa yhtä aineiston muuttujaa. Muuttujaan liittyy seuraavia metatietoja.

Taulukko 25 aineistossa olevaa muuttujaa kuvailevat tiedot

Nro	Metatieto	Määrä	Selitys	Esimerkki	Tärkeys
M1	muuttuja	1	muuttujan nimi (= kentän nimi) aineistossa	HTIKA	MUST (a)
M2	selite	1	kuvaus siitä, mitä muuttujaan on talletettu	vanhemman ikä kokonaisuina vuosina vuoden alussa	MUST
M3	tyyppi	1	muuttujan mahdollinen arvojoukko ja esitystapa	luku välillä 0-120 / päivämäärä muodossa vvvv-kk-pp	MUST osin (a)
M4	koodisto	0-1	koodatun muuttujan arvojen merkitys, koodiston nimi (jolla arvot voi etsiä metatietojen koodilistasta) tai viittaus toisaalla määriteltyyn koodistoon	1=nainen, 2=mies / kunta- ja alueluokitus	MUST
M5	muutostieto	0-*	useammassa erässä tai jatkuvasti kerätyissä aineistoissa: muuttujan keräystavassa, tulokinnassa tai merkinnässä tapahtuneet muutokset liitettyä ajankohtaan, jolloin käytäntö on vaihtunut	2001 alkaen koodit naimaton, leski ja eronnut yhdistetty koodiksi naimaton	MUST

M6	käsittely	0-*	miten muuttujan arvo on käsitelty ennen tallettamista / luovuttamista, oletus: ei mitenkään	karkeistettu 10 vuoden tarkkuudelle	MUST
M7	puuttuminen	0-*	miten puuttuva tieto on merkitty aineistossa; jos useampia puuttumisen merkitsimiä, myös selitys, millaista puuttumista mikin tarkoittaa	NULL	MUST
M8	nimi	1	ihmisten ymmärrettäväksi tarkoitettu muuttujan lyhyt nimi	ikä	SHOULD
M9	lähde	1	muuttujan alkuperä, myös kysymys / ohjeet, joista tieto on tuloksena / kaava, jolla johdettu muuttuja on päätelty	lasten määrä on saatu laskemalla yhteen kaikki ne henkilöt, joiden isäksi / äidiksi henkilö on merkitty	SHOULD
M10	aikajakso	1-*	pitkittäisaineistoissa, miltä ajalta muuttuja on kerätty	vuodet 1998-2006	SHOULD
M11	kattavuus	1	kuinka suurella osalla riveistä tieto ei puutu; ks. myös "puuttuva" koodiston tiedoissa	93%	SHOULD (a)
M12	käyttöoikeudet	1-*	kriteerit, joilla muuttuja luovutetaan aineiston mukana	ei luovuteta	SHOULD
M13	turvaluokitus	0-*	maininta, onko tieto arkaluontoinen, henkilötieto ja/tai henkilötunnus (oletus: julkinen tieto)	tunnistetieto	SHOULD
M14	ryhmä	0-1	toisiinsa liittyville muuttujille yhteinen nimitys	perushenkilötiedot	MAY
M15	pakollisuus	1	onko tieto ollut pakollinen tietoja kerättyessä	ei	SHOULD
M16	tunnusluvut	0-1	numeerisen muuttujan tilastolliset tunnusluvut (minimi, maksimi, keskiarvo, mediaani, keskihajonta)	min=0, max=98, avg=45.5, med=43.7, dev=31.1	MAY (a)
M17	frekvenssit	0-1	koodatun muuttujan eri koodien yleisyys	nainen=51.2%, mies=48.4%	MAY (a)
M18	käyttötarkoitus	1-*	muuttujan pääasiallinen käyttötarkoitus	taustatieto tilastoinnissa	MAY
M19	riippuvuudet	0-*	yhteydet toisiin muuttujiin, myös tarkistussäännöt	lasten määrä = 0 jos ikä < 10, vastaa v.2006 muuttujaa HTIKA06	MAY

## Koodien tiedot

Koodit ovat distinktiivisiä arvoja, jotka tietty muuttuja voi saada (se, mitä koodistoa muuttuja käyttää, on muuttujan metatieto). Jokaisella mahdollisella arvolla on omat metatietonsa.

Taulukko 26 muuttujan koodaukseen käytettyä koodia kuvailevat tiedot

Nro	Metatieto	Määrä	Selitys	Esimerkki	Tärkeys
-----	-----------	-------	---------	-----------	---------

K1	koodisto	1	koodiston nimi, johon koodi kuuluu, käytetään myös viittaamaan koodistoon muuttujien kuvauksissa	sosioekonominen asema	MUST
K2	koodi	1	koodistoa käyttävän muuttujan arvo silloin, kun halutaan viitata tähän koodiin	9	MUST
K3	selite	1	koodin merkitys ihmisen ymmärrettäväksi tarkoitettussa muodossa	työskentelee ulkomailla	MUST
K4	määrittely	1	tarkka kuvaus kriteereistä, joilla koodi valitaan	henkilön vakinainen osoite on ollut vuoden aikana ulkomailla	SHOULD
K5	viitteet	0-*	taustamateriaali / autoritatiivinen koodiston määrittely	esim. linkki asiasanastoon	SHOULD
K6	aikajakso	1-*	aika, jolloin koodia on käytetty tietojen merkinnässä	2002-	SHOULD

## Liite 7. MIDRAS-metatiedon vastaavuus DDI3-elementteihin

Tässä liitteessä esitellään, miten MIDRAS-järjestelmän metatiedon vaatimusmäärittelyn eri tiedot ilmaistaan DDI3.1-standardin mukaisesti.

DDI3.1-määrittelyllä on monimutkainen rakenne, eikä kaikissa tapauksissa ole yksiselitteistä, mihin osaan koko aineiston määrittelyssä mikin elementti tai attribuutti kuuluu. Kun elementti tai attribuutti voi sijaita useammassa kohdassa DDI3.1-määrittelyä, on erikseen merkitty, minkä elementin sisällä annetun elementin tai attribuutin on oltava. Alla olevassa taulukossa on esitelty, mitä DDI3-elementtien merkinnässä käytetyt merkintätavat tarkoittavat.

Taulukko 27 vastaavuustaulukossa käytettyjen merkintöjen selitys

Merkintä	Merkitys	Esimerkki	Esimerkin merkitys DDI:ssa
sana tai sanaryhmä	Elementin tai attribuutin nimi	version rationale	<VersionRationale> ... </VersionRationale>
nimi1 / nimi2	Elementti toisen sisällä tai tietyn elementin attribuutti	DDI instance / version date	<DDIInstance versionDate="...">
		study unit / embargo	<StudyUnit> <Embargo> ... </Embargo> </StudyUnit>
elementti1 -> elementti2	Elementissä on viittaus (IDs.) toiseen elementtiin	concept reference -> concept / description	<ConceptReference> <ID>XYZ123</ID> </ConceptReference> ... <Concept id="XYZ123"> <Description> ... </Description> </Concept>
kattaa:	DDI-elementti tai attribuutti sisältää osan MIDRAS-tiedosta	kattaa: DDI instance / is published	isPublished-attribuutti kertoo, onko metatiedot julkistettu, mutta ei yksityiskohtaisempaa julkisuusmäärittelyä
sisältyy:	DDI-elementti tai attribuutti voi sisältää muitakin tietoja	sisältyy: study unit / embargo	Embargo-elementti kertoo kaikenlaiset käyttörajoitukset mukaan lukien arkaluonteisuusrajoitukset
pääteltävissä:	MIDRAS-tieto on muodostettavissa DDI-elementin tai -attribuutin tai useamman sisällöstä säännön perusteella	pääteltävissä: total responses	TotalResponses-elementti sisältää niiden havaintojen määrän, joista muuttuja on määritelty; tästä voi laskea, kuinka suuressa osassa havainnoista muuttuja on määritelty
ilmaistavissa:	MIDRAS-tiedon ilmaisuun on keinot käyttämällä DDI-elementtiä tietyllä tavalla	ilmaistavissa: response domain / missing value	ResponseDomain-elementin missingValue-attribuutin voi asettaa tyhjäksi, jos puuttuvaa arvoa ei voi olla

Alla olevassa taulukossa on nimetty kukin MIDRAS-metatiedon osa ja sitä vastaava DDI3.1-elementti tai attribuutti. Taulukko kertoo, miten erilaiset MIDRAS-järjestelmässä vaaditut aineistojen kuvailut esitetään DDI3.1-standardissa.

Taulukko 28 MIDRAS- ja DDI3-elementtien vastaavuus

Nro	MIDRAS	DDI3
-----	--------	------

MT1	julkisuus	kattaa: DDI instance / is published
MT2	standardi	DDI instance / xmlns
MT3	muutospäivä	DDI instance / version date
MT4	muutostieto	DDI instance / version rationale
MT5	versio	DDI instance / version
MT6	lähde	DDI instance / version responsibility
MT7	kieli	DDI instance / xml:lang
A1	nimi	title
A2	ylläpito-organisaatio	publisher
A3	yhteystieto	individual
		organization
A4	havaintoyksikkö	analysis unit
A5	kohde	coverage
A6	otos	sisältyy: sampling procedure
A7	koko	case quantity
A8	muuttujamäärä	logical record / variable quantity
A9	kuvaus	abstract
A10	aikajakso	temporal coverage
A11	päivitysväli	data collection frequency
A12	muutostieto	group / version rationale
		data collection / version rationale
A13	käsittely	processing event
A14	käyttöoikeudet	sisältyy: study unit / embargo
A15	tilausprosessi	access
A16	luovutusmuoto	gross file structure
		format
A17	sarja	series name
A18	asiasanat	topical coverage / keyword
A19	lähde	data source
		data collector organization reference
A20	keruumenetelmä	methodology
A21	luontipäivä	pääteltävissä: study unit / version + study unit / version date
A22	muutospäivä	study unit / version date
A23	versio	study unit / version
A24	käyttötarkoitus	purpose
A25	viitteet	study unit / other material / relationship
A26	julkisuus	sisältyy: access
A27	kieli	language of data
A28	sijainti	data file identification

A29	tunnus	study unit / id
M1	muuttuja	variable / label
M2	selite	variable / concept reference -> concept / description
		variable / question reference -> question item / question text
M3	tyyppi	representation
M4	koodisto	code scheme reference
M5	muutostieto	variable / version date
		variable / version rationale
M6	käsittely	processing event
M7	puuttuminen	excluded missing category reference -> category
		response domain / missing value
M8	nimi	variable name
M9	lähde	representation / coding instruction reference -> coding / generation instruction
M10	aikajakso	variable / version date
M11	kattavuus	pääteltävissä: total responses
M12	käyttöoikeudet	sisältyy: variable / embargo
M13	turvaluokitus	sisältyy: variable / embargo
M14	ryhmä	variable group name
M15	pakollisuus	ilmaistavissa: response domain / missing value
		ilmaistavissa: variable group / group type coded
M16	tunnusluvut	summary statistics
M17	frekvenssit	category statistics
M18	käyttötarkoitus	kattaa: representation / role
		question intent
M19	riippuvuudet	data relationship
K1	koodisto	code scheme / code scheme name
K2	koodi	code scheme / code / value
K3	selite	code scheme / code / category reference -> category / category name
K4	määrittely	category / description
K5	viitteet	category scheme reference
K6	aikajakso	pääteltävissä: category / version date