AVOIN TIEDE JA TUTKIMUS

TUTKIMUSAINEISTOJEN TIEDOSTOMUODOT JA PITKÄAIKAISSÄILYTYSKELPOISUUS

SELVITYKSEN LOPPURAPORTTI

Julkaisu Tutkimusaineistojen tiedostomuodot ja pitkäaikaissäilytyskelpoisuus	
Julkaisija Avoin tiede ja tutkimus -hanke	**Julkaisuajankohta** 10.2.2017
Tekijä Tutkimus-PAS -työryhmä	
Lisenssi	
Julkaisun jakelun http://urn.fi/URN:NBN:fi-fe2017121855905 PDF-tiedosto ladattavissa sivuilla avointiede.fi/keskeiset-julkaisut	
Yhteystiedot http://avointiede.fi avointiede@postit.csc.fi	

SISÄLTÖ

1 TIIVISTELMÄ

Avoin tiede ja tutkimus -hanke[1] on opetus- ja kulttuuriministeriön vuonna 2014 käynnistämä hanke tiedon saatavuuden ja avoimen tieteen edistämiseksi. Tärkeä osa-alue hankkeessa on tutkimuksen tuotosten pitkäaikaissäilytys (PAS) ja pitkäaikaisen saatavuuden turvaaminen. Tähän tarkoitukseen kehitetään kestäviä toimintamalleja, joilla voidaan varmistaa aineistojen käyttökelpoisuus jopa useiden kymmenien tulevien vuosien aikana. Tutkimusaineistojen PAS-kokonaisuus (TPAS-kokonaisuus) sisältää toimintamalleja tukevat palvelut ja järjestelmät, joilla toteutetaan aineistojen pitkäaikaissäilytys sekä sen rajapinnat ja käyttöliittymät.

Tämä selvitys on osa TPAS-kokonaisuuden ja säilytyksen toteuttavan PAS-ratkaisun suunnittelua. Siinä keskitytään tutkimusaineistojen tiedostomuotoihin, joiden ymmärrettävyys, levinneisyys ja ohjelmistotuki ovat tärkeitä aineistojen uudelleenkäytön kannalta. Asiaan perehdyttiin kansainvälisten lähteiden avulla ja haastattelemalla suomalaisia tutkijoita. Lisäksi laadittiin alustavat vaatimukset aineistojen hyväksymiseksi pitkäaikaissäilytykseen.

1.1 Esimerkkiaineistot

Tutkimusaineistot koostuvat lähes aina useista toisiinsa liittyvistä tiedostoista. Aineisto voi sisältää esimerkiksi mittalaitteesta saatua raakadataa, mittalaitteiden asetukset kertovia metatietoja, koetilanteen kuvauksen ja tutkimuksen tulokset kuvaavan julkaisun. Olennaista on, että mukana olevat tiedostot muodostavat aineistoa käyttävälle tutkijalle ymmärrettävän kokonaisuuden.

Selvityksessä haastatelluilta tutkijoilta saatiin analysoitavaksi alla olevassa taulukossa esitetyt yksitoista esimerkkiaineistoa, joita tarkasteltiin pitkäaikaissäilytyksen näkökulmasta. Otos ei kata kaikkia tieteenaloja, mutta antaa hyvän yleiskuvan aineistotyypeistä ja tiedostomuodoista.

Lyhenne	Luoja / omistaja	Tieteenala	Aineiston kuvaus
1000Gen	Kansainvälinen 1000 genomia -projekti	Bio- ja terveystieteet	Kansainvälisenä yhteistyönä kerättyjä ihmisen perimän geenisekvenssejä
Aivokuvat	Aalto-yliopisto, Aivo ja Mieli -laboratorio	Lääketieteellinen tekniikka	Magneettikuvia aivoista koehenkilön katsoessa elokuvaa
ERNE	Turun yliopisto, Avaruustutkimus-laboratorio	Luonnontieteet, avaruustutkimus	ERNE-hiukkashavainnoijan mittaustuloksia kosmisesta säteilystä
FIRE	Helsingin yliopisto, Seismologian instituutti	Ympäristötieteet, seismologia	Seismologisia mittauksia Suomen maaperästä
FSD	Yhteiskuntatieteellinen tietoarkisto FSD	Yhteiskunta- ja humanistiset tieteet	Kyselyt suomalaisten median käytöstä ja suhteesta kulttuuriperintöönsä
Kiteet	Aalto-yliopisto, Biotalousinfrastruktuuri	Luonnontieteet, biokemia	Mittaustuloksia liittyen kiteiden muodostumiseen pehmeissä aineissa

[1] http://avointiede.fi

Lyhenne	Luoja / omistaja	Tieteenala	Aineiston kuvaus
MAXIV	MAX IV -laboratorio, Ruotsi ja Oulun yliopisto	Luonnontieteet, materiaalifysiikka	Röntgenmikroskopiaan liittyvä esimerkkitiedosto
Planck	Euroopan avaruusjärjestö ESA	Luonnontieteet, avaruustutkimus	Planck-avaruustutkimusaseman kosmisen taustasäteilyn mittauksia
RITU	Jyväskylän yliopisto, Kiihdytinlaboratorio	Luonnontieteet, hiukkasfysiikka	Kiihdytinlaboratorion RITU-rekyyliseparaattorin mittaustuloksia
SMEAR	Helsingin yliopisto, SMEAR-tutkimusasemat	Ympäristötieteet, ilmakehätieteet	Tietokanta useiden eri mitta-laitteiden ja havaintoasemien mittaustuloksista
Suomi24	CSC ja Kotimaisten kielten keskus	Yhteiskunta-tieteet, Kielitiede	Suomi24-keskustelufoorumin viestejä kielitieteellisesti jäsennettynä

Lähes jokainen aineisto koostui useammassa eri tiedostomuodossa olevista tiedostoista. Yhteensä esimerkkiaineistoissa esiintyi 26 eri tiedostomuotoa, joista puolet on jo hyväksytty Kansallisen digitaalisen kirjaston pitkäaikaissäilytyksessä (KDK-PAS) säilytys- tai siirtokelpoisiksi. Lopuistakin valtaosa oli avoimia ja dokumentoituja.

Suurin osa esimerkkiaineistojen tiedostomuodoista voidaan hyväksyä pitkäaikaissäilytykseen, kun niille on laadittu teknisten metatietojen määritykset. Lisäksi aineistot tulee dokumentoida huolellisesti. Tutkimusaineiston kaikki osatekijät paketoidaan. Paketointi tässä yhteydessä tarkoittaa ennen kaikkea aineistokokonaisuuden eri osien roolin ja keskinäisten suhteiden sekä niiden metatietojen esitystä standardilla tavalla.

Esimerkkiaineistojen lisäksi haastatteluissa ja selvityksessä kartoitettiin muita tutkimusaineistoissa yleisesti käytettyjä tiedostomuotoja sekä tietokantoja. Tiedostomuotojen kirjo on kulttuuriaineistoja suurempi, ja monet muodot ovat tiedealakohtaisia. Tietokantamuotoiset aineistot ovat säilytyksen kannalta erityinen haaste. Lisäksi aineistojen ymmärtäminen edellyttää usein alan asiantuntemusta. Kansainvälinen yhteistyö ohjaa kuitenkin tutkijoita käyttämään yhä paremmin yhteensopivia ja dokumentoituja tiedostomuotoja, mikä helpottaa pitkäaikaissäilytystä.

1.2 Aineistojen hyväksyminen pitkäaikaissäilytykseen

Tavoitteena on tehdä aineistojen toimittamisesta säilytykseen helppoa ja kätevää, jotta aineistot saadaan mahdollisimman laajasti ja nopeasti uudelleenkäytettäväksi. Lähtökohdaksi otettiin Kansallisen digitaalisen kirjaston pitkäaikaissäilytyksen (KDK-PAS) vaatimukset, joita muokattiin tutkimusaineistojen erityispiirteet huomioiden.

Pitkäaikaissäilytys varmistaa aineiston ymmärrettävyyden säilymisen erittäin pitkällä aikavälillä, yli teknologian, tutkimuskäytäntöjen ja muiden merkittävien muutosten. Se asettaa verraten tiukkoja vaatimuksia aineiston tiedostomuodoille ja metatiedoille. Pitkäaikaissäilytettävien tutkimusaineistojen vaatimukset vastaavat KDK-PAS:ssa kulttuuriaineistoille asetettuja vaatimuksia.

Esimerkiksi tilanteessa, jossa päätöstä aineiston varsinaisesta pitkäaikaissäilytyksestä halutaan vielä harkita, voidaan sen saatavuus varmistaa useiksi vuosiksi eheyden säilytyksellä. Tällaisesta säilytyksestä käytetään tässä dokumentissa nimitystä keskipitkä säilytys. Siinä erityisesti tiedostomuodoille asetettavia vaatimuksia on lievennetty varsinaiseen pitkäaikaissäilytykseen verrattuna. Aineisto ja sen osat tulee kuitenkin myös keskipitkässä

säilytyksessä olla kuvailtu siten, että kokonaisuus on muille tutkijoille käyttökelpoinen. Lisäksi käyttöoikeustiedot tulee olla ilmoitettu.

Vaatimukset aineiston hyväksymiseksi säilytykseen:

1. Aineistokokonaisuus on muille tutkijoille käyttökelpoinen. (pakollinen)
2. Kokonaisuuteen kuuluvat tiedostot ja niiden suhteet on kuvattu PAS-ratkaisun määritysten mukaisesti. (pakollinen)
3. Tiedostot ovat joissakin PAS-ratkaisussa hyväksytyistä säilytys- tai siirtokelpoisista tiedostomuodoista. (pakollinen, keskipitkässä säilytyksessä suositeltava)
4. Aineiston käyttöoikeustiedot on ilmoitettu. (pakollinen)
5. Aineiston lisenssi on voimassaolevien avoimen tieteen suositusten mukainen. (suositeltava)
6. Aineisto on kuvailtu metatietomääritysten mukaisesti. (pakollinen)

Vaatimukset keskipitkään säilytykseen vastaanotettaville tiedostomuodoille, joita ei ole hyväksytty säilytys- tai siirtokelpoisiksi:

1. Tiedostomuoto on tuettu vähintään yhdessä yleisesti saatavilla olevassa ohjelmassa (pakollinen)
2. Tiedostomuodon rakenne on dokumentoitu (suositeltava)
3. Tiedostomuoto on alalla laajasti käytetty (suositeltava)
4. Tiedostomuoto on riippumattoman organisaation tai alan tutkimusyhteisön standardoima (suositeltava)

Säilytys- ja siirtokelpoisiksi hyväksytyt tiedostomuodot täyttävät luonnollisesti kaikki pakolliset vaatimukset ja harvoja poikkeuksia lukuun ottamatta myös suositukset.

Vaatimukset täyttävä aineisto voidaan siirtää suoraan pitkäaikaissäilytykseen. Vaihtoehtoisesti aineisto voidaan julkaista ensin keskipitkän säilytyksen palvelussa, jolloin mahdollisesti tarvittava tiedostomuotojen hyväksyntäprosessi ei viivytä julkaisua ja aineiston siirtämisestä varsinaiseen pitkäaikaissäilytykseen päätetään myöhemmin.

1.3 KDK:n pitkäaikaissäilytyksestä tutkimusaineistojen säilytykseen

KDK:n pitkäaikaissäilytyksen määritykset muodostavat hyvän pohjan tutkimusaineistojen pitkäaikaissäilytykselle. Olemassa olevia määrityksiä voidaan laajentaa kattamaan uudet aineistotyypit ja tiedostomuodot, sekä tehdä tarvittavat muutokset prosesseihin ja vastuualueisiin.

Myös KDK:ssa laadittu aineistojen paketointimalli soveltuu tutkimusaineistoille. Aineistojen kuvailu- ja paketointipalveluiden helppokäyttöisyyteen on syytä erityisesti panostaa.

KDK:n aineistoista vastaa yleensä museo, kirjasto tai arkisto, jolla on lakisääteinen säilytystehtävä. Tutkimusaineistot on tyypillisesti laadittu hankkeissa, joilla on päättymispäivä eikä pitkäaikaista vastuuta aineistojen säilytyksestä. Säilytykseen siirrosta vastaava organisaatio saattaa olla tutkimusinfrastruktuuri, joka hallinnoi tietyn tieteenalan aineistoja yliopistorajat ylittäen ja jolla on hyvät valmiudet dokumentoida aineistot yhtenäisesti.

Tutkimusaineistojen pitkäaikaissäilytys on myös kansainvälisesti melko varhaisessa vaiheessa. Suurin osa organisaatioista keskittyy keskipitkään säilytykseen, eikä kattavia listoja suositeltavista tiedostomuodoista tai määrityksiä ymmärrettävyyden kannalta olennaisista metatiedoista ole vielä laadittu. Suomella on kansallisen PAS-ratkaisun myötä mahdollisuus olla edelläkävijä ja haluttu kumppani myös kansainvälisten tutkimusaineistojen säilytyksen osalta.

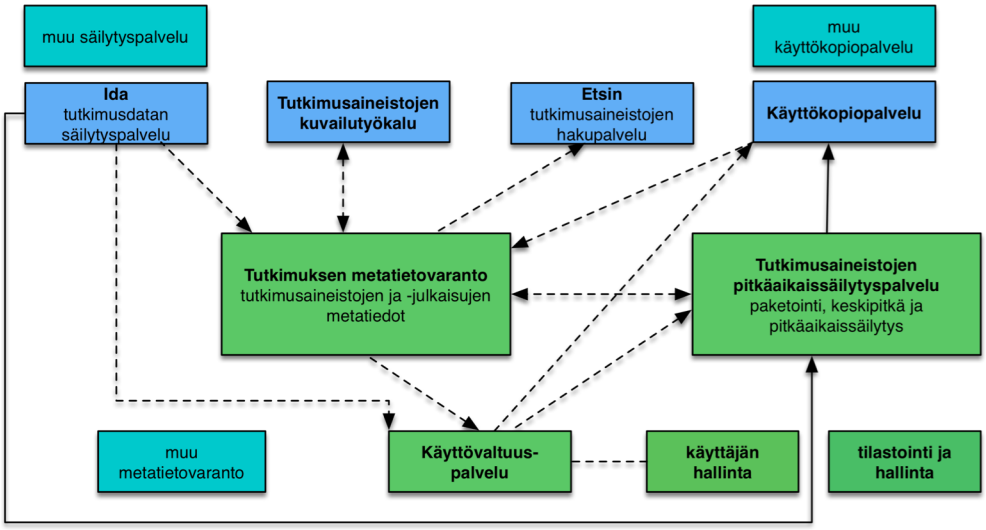
2 JOHDANTO

Avoin tiede ja tutkimus -hanke[2] on opetus- ja kulttuuriministeriön vuonna 2014 käynnistämä hanke tiedon saatavuuden ja avoimen tieteen edistämiseksi. Tavoitteena on, että vuoteen 2017 mennessä Suomi nousee yhdeksi johtavista maista tieteen ja tutkimuksen avoimuudessa ja että avoimen tieteen mahdollisuudet hyödynnetään laajasti yhteiskunnassa. Lisäksi tavoitteena on edistää tieteen ja tutkimuksen luotettavuutta, tukea avoimen tieteen ja tutkimuksen toimintatavan sisäistämistä tutkijayhteisössä sekä lisätä tutkimuksen ja tieteen yhteiskunnallista ja sosiaalista vaikuttavuutta.

Tärkeä osa-alue hankkeessa on tutkimuksen tuotosten pitkäaikaissäilytys (PAS) ja pitkäaikaisen saatavuuden turvaaminen. Digitaalisen informaation pitkäaikaisen saatavuuden turvaaminen on sellaisten kestävien toimintamallien rakentamista, joilla voidaan varmistaa aineistojen käyttökelpoisuus jopa useiden kymmenien tulevien vuosien aikana, yli teknologisten ja tutkimuskäytäntöjen muutosten. Kansainvälinen, usein tieteenalakohtainen yhteentoimivuus tulee varmistaa sopivalla yhteistyöllä, sopimuksilla ja semanttisen tason tietomalleilla.

Tutkimusaineistojen hyväksyminen pitkäaikaissäilytettäväksi ja tarjoaminen sen puitteissa muiden tutkijoiden saataville on monitasoinen prosessi, johon liittyy sekä vastuu-, käyttöoikeus- että teknisiä kysymyksiä. Teknisellä tasolla varmistetaan, että aineisto on eheä, ymmärrettävä ja soveltuu uudelleenkäytettäväksi. Tutkimusaineistojen pitkäaikaissäilytys-kokonaisuus eli TPAS-kokonaisuus (Kuva 1) sisältää palvelut ja järjestelmät, joilla toteutetaan säilytyksen vaatimat toiminnallisuudet ja tarjotaan niiden käyttöön tarvittavat rajapinnat sekä käyttöliittymät. Alustavien suunnitelmien mukaan sekä KDK:n kulttuuriaineistojen että tutkimusaineistojen säilytykseen soveltuu tekniseltä alustaltaan yhteinen PAS-ratkaisu.

Tutkimusaineistot koostuvat lähes aina useista toisiinsa liittyvistä tiedostoista. Aineisto voi



Kuva 1: TPAS-kokonaisuus

sisältää esimerkiksi mittalaitteesta saatua raakadataa, analyysin tuloksena saatuja datatiedostoja, mittalaitteiden asetukset kertovia metatietoja, koetilanteen kuvauksen ja tutkimuksen tulokset kuvaavan julkaisun. Olennaista on, että mukana olevat tiedostot muodostavat aineistoa uudelleenkäyttävälle tutkijalle ymmärrettävän kokonaisuuden.

Tässä selvityksessä keskitytään erityisesti tutkimusaineistojen tiedostomuotoihin, jotka ovat yksi keskeisistä seikoista uudelleenkäytön kannalta. Yleisesti käytettyjä tiedostomuotoja selvitettiin sekä kansainvälisten lähteiden avulla että haastattelemalla Suomessa aineistojen

[2] http://avointiede.fi

parissa työskenteleviä tutkijoita. Osana selvitystä tarkasteltiin myös käytännönläheisesti yhtätoista ajankohtaista tutkimusaineistoa.

Selvityksen alussa esitellään työmenetelmät sekä analysoitavaksi saadut esimerkkiaineistot. Tiedostomuotojen analyysi -luvussa käsitellään sekä esimerkkiaineistojen tiedostomuotoja että muita eri tieteenaloilla yleisesti käytössä olevia tiedostomuotoja ja tietokantoja, sekä arvioidaan niiden ominaisuuksia aineistojen uudelleenkäytön kannalta. Seuraavassa luvussa luetellaan alustavat vaatimukset aineistojen hyväksymiseksi pitkäaikaissäilytykseen, kuvataan aineistojen hyväksymisprosessi sekä evaluoidaan miltä osin esimerkkiaineistot täyttävät vaatimukset. Lopuksi esitetään johtopäätökset sekä tarpeet jatkotöiksi. Selvityksen liitteinä ovat lista haastatelluista henkilöt, haastattelukysymykset ja yksityiskohtaiset taulukot tiedostomuotojen analyysista.

3 TYÖMENETELMÄT

Aihetta lähestyttiin kahdesta näkökulmasta. Ajantasaiseen kansainväliseen tietoon perehdyttiin tiedostomuotoja käsittelevien dokumenttien ja Internet-sivustojen kautta, sekä olemalla yhteydessä muutamiin yhteistyökumppaneihin. Toisena peruspilarina olivat suomalaisten tutkijoiden ja tutkimusryhmien haastattelut, joiden avulla kerättiin tietoa eri tutkimusalojen aineistoista sekä kotimaisista tarpeista. Tutkimusryhmiltä pyydettiin myös esimerkkiaineistoja, joihin tutustuttiin ja joita analysoitiin yksityiskohtaisesti tiedostotasolla.

Haastatellut henkilöt sekä esimerkkiaineistot valittiin siten, että niiden pohjalta saatiin riittävän kattava yleiskuva erilaisista aineistotyypeistä ja niiden tiedostomuodoista pitkäaikaissäilytyksen ja -saatavuuden näkökulmasta. Otos ei kata kaikkia tieteenaloja, mutta antaa hyvän pohjan yleisten aineistoihin liittyvien vaatimusten ja prosessien laatimiseen. Tarkennuksia, lisäyksiä ja muutoksia voidaan tehdä tieteenalakohtaisesti myöhemmin.

Tiedostomuotojen soveltuvuutta arvioitaessa hyödynnettiin myös eri maissa toimivien, aineistoja pitkäaikaissäilyttävien organisaatioiden julkaisemia listoja sekä kuvauksia heidän hyväksymistään ja suosittelemistaan tiedostomuodoista. Tiedostomuotojen ohjelmistotuen osalta tukeuduttiin julkisesti saatavilla oleviin tietoihin, eikä ohjelmistoja kokeiltu muutamaa poikkeusta lukuun ottamatta.

Projektin tavoitteet ja linjaukset laativat Tutkimus-PAS-hankkeen projektipäällikkö Esa-Pekka Keskitalo Kansalliskirjastosta sekä ATT-hankkeen pääsihteeri Pirjo-Leena Forsström. Käytännön työstä ja raportin kirjoittamisesta päävastuussa oli pitkäaikaissäilytykseen erikoistunut konsultti Arto Teräs. Monissa haastatteluissa oli mukana myös Juha Törnroos CSC:ltä. Vaatimukset aineistojen hyväksymiseksi pitkäaikaissäilytykseen laadittiin selvityksessä kerätyn tiedon pohjalta yhteistyönä Tutkimus-PAS-projektiryhmässä. Tietolähteinä keskeisessä roolissa olivat haastatellut henkilöt (Liite A), joita projekti lämpimästi kiittää. Heillä oli mahdollisuus kommentoida raporttia ennen sen julkaisua.

4 KANSAINVÄLINEN KATSAUS

Tutkimushankkeissa tehdään yhä laajemmin kansainvälistä yhteistyötä ja tutkimusaineistoilla on kansainvälistä mielenkiintoa. Varsinkin laajojen kansainvälisten hankkeiden aineistoja säilytetään usein keskitetysti tietopankeissa, joita tutkijat käyttävät tietolähteinä ja joihin useiden eri maiden tutkijat lähettävät aineistoa. Kansainvälisen yhteistyön helpottamiseksi on tärkeää huomioida maailmalla tehdyt ratkaisut ja käytännöt kansallista säilytystä suunniteltaessa. Huolehtimalla tiedostomuotojen sekä metatietojen yhteensopivuudesta kansainvälisten toimijoiden kanssa suomalaisten tutkijoiden tietojen vaihto ulkomaisten kollegoiden sekä organisaatioiden kanssa helpottuu. Suosittujen tiedostomuotojen käsittelyyn on myös saatavilla valmiita validointi- ja muita työkaluja, mikä laskee säilytyksen kustannuksia.

Avoin tiede -hankkeessa on jo aiemmin julkaistu tutkimusdatan pitkäaikaissäilytyksen kansainvälinen katsaus, jossa keskityttiin palveluratkaisuihin ja prosesseihin sekä niitä tukeviin yhteistyön ja hallinnon järjestelyihin neljässä eri maissa sijaitsevassa organisaatiossa [ATT_KVKatsaus]. Tässä selvityksessä paneuduttiin tarkemmin kansainvälisesti tehtyihin valintoihin ja suosituksiin tiedostomuotojen osalta. Tarkastelussa olivat mukana Australian kansallisarkisto, ranskalainen CINES, hollantilainen DANS, isobritannialainen UKDA, kanadalainen Library and Archives Canada (LAC) ja yhdysvaltalainen Library of Congress.

Ulkomaisten organisaatioiden tekemät valinnat säilytettäviksi ja siirtokelpoisiksi tiedostomuodoiksi ovat monilta osin yhteneviä KDK-hankkeessa tehtyjen valintojen kanssa [KDK_Tiedostomuodot]. KDK:n määritys on tiedostomuotojen versioiden ja niihin liittyvien metatietojen suhteen täsmällisempi kuin suurin osa kansainvälisistä ohjeista ja suosituksista. Toisaalta siitä puuttuu useita tutkimusaineistoissa esiintyviä tiedostomuotoja, jotka kansainvälisissä suosituksissa on huomioitu.

Australian kansallisarkiston ja UKDA:n dokumentit [NAA_Formats] [UKDA_Formats] ovat yksinkertaisia listoja säilytykseen hyväksytyistä ja suositeltavista tiedostomuodoista, ilman tarkempia ohjeita niiden versioista tai metatiedoista. CINES mainitsee tiedostomuotojen versiot, ja kyseessä on samalla lista muodoista, jotka organisaation tarjoama säilytyspalvelu pystyy vastaanoton yhteydessä validoimaan [CINES_Formats].

DANSin Preferred Formats -dokumentti [DANS_Formats] antaa hyväksyttyjen muotojen listan lisäksi tarkempia ohjeita kuhunkin tiedostomuotoon tai kategoriaan liittyen. Mukana on useita KDK:ssa toistaiseksi käsittelemättömiä aineistotyyppejä, mm. paikkatieto-, CAD-, ja 3D-aineistot sekä tietokannat. Ohjeistuksen tarkkuus ei kuitenkaan ulotu tiedostomuotojen versioiden tai niihin liittyvien metatietojen tasolle.

Library and Archives Canada on julkaissut kaksikin dokumenttia, joista vanhemmassa vuoden 2010 versiossa tiedostomuotoja on arvioitu varsin perusteellisesti [LAC_Formats_2010]. KDK:n säilytys- ja siirtokelpoisten tiedostomuotojen arviointikriteerit perustuvat tähän dokumenttiin. Uudempi vuonna 2015 julkaistu lista [LAC_Formats] on hiukan suppeampi, mutta sisältää varsin tarkat suositeltujen muotojen versiotiedot sekä yleiset perustelut niiden valinnasta. Mukana on useita KDK:ssa toistaiseksi käsittelemättömiä aineistotyyppejä.

Library of Congressin dokumentti [LoC_Statement] on lähestymistavaltaan erilainen: siinä ei pyritä kattavaan listaan yksittäisistä säilytykseen hyväksyttävistä tiedostomuodoista, vaan antamaan yleisluontoisempia suosituksia sekä luomaan kategorioittain karkea paremmuusjärjestys. Joitakin tiedostomuotoja on mainittu nimeltä, mutta mukana on myös yleisempiä luokkia kuten "merkintäkielet", "julkisesti dokumentoidut tiedostomuodot" ja "laajasti käytetyt valmistajakohtaiset muodot". Itse tiedostomuotojen lisäksi dokumentti sisältää vaatimuksia ja ohjeita metatietoihin sekä aineiston sisältöön liittyen.

Erityishuomion ansaitsee samaisen Library of Congressin Sustainability of Digital Formats -sivusto [LoC_Formats]. Sivuilta löytyy sekä yleisiä ohjeita tiedostomuotojen valintaan liittyen että yksityiskohtaiset kuvaukset monista suosituista muodoista. Lista ei kata lukumääräisesti yhtä suurta määrää muotoja kuin PRONOM-tiedostomuotokirjasto

[PRONOM], mutta on täsmällisten ja asiantuntevien kuvaustensa ansiosta kokonaisuutena hyödyllisempi tietolähde.

Kansainväliset suositukset huomioitiin tarkasteltaessa esimerkkiaineistoja sekä eri tieteenaloilla käytettyjä tiedostomuotoja. Tehdyt huomiot on kirjattu osaksi selvityksen tekstiä tiedostomuotojen kuvausten yhteyteen.

5 ESIMERKKIAINEISTOT

Analysoitavaksi saatiin yksitoista esimerkkiaineistoa, joiden perustiedot on esitetty alla olevassa taulukossa:

Lyhenne	Luoja / omistaja	Tieteenala	Aineiston kuvaus ja analysoitavaksi valittu osa
1000Gen	Kansainvälinen 1000 genomia -projekti	Bio- ja terveystieteet	Kansainvälisenä yhteistyönä kerätty 1000 ihmisen perimän sekvenssit sisältävä aineisto. Aineisto on vapaasti ladattavissa verkosta. [1000GENOMES] Analysoitavaksi valittiin yhden koehenkilön (HG00180) perimän sekvenssi yleisimmin käytetyissä tiedostomuodoissa.
Aivokuvat	Aalto-yliopisto, Neurotieteen ja lääketieteellisen tekniikan laitos, Aivo ja Mieli -laboratorio	Bio- ja terveys-tieteet, Lääke-tieteellinen tekniikka	Aineisto koostuu magneettikuvauksella (MRI) saaduista kuvasarjoista, jotka kuvaavat aivojen toimintaa koehenkilön katsoessa elokuvaa. [AALTO] Analysoitavaksi valittiin kolmen koehenkilön kuvasarjat sekä niihin liittyvät oheistiedostot, mm. tutkimuksessa käytetty elokuva. Kyseessä oli osa vuonna 2015 tehdyssä pilotissa käytetystä aineistosta. [PAS_Pilotit_2015]
ERNE	Turun yliopisto, Fysiikan ja tähtitieteen laitos, Avaruustutkimus-laboratorio	Luonnontieteet ja tekniikka, Avaruus-tutkimus	Aineisto sisältää ERNE-hiukkas-havainnoijan mittaustuloksia laitteeseen osuvien kosmisen säteilyn hiukkasien energioista. [ERNE] Analysoitavana oli otos mittaustuloksista, jonka tutkijat olivat jo aiemmin valinneet vuonna 2015 tehtyä pitkäaikaissäilytyksen pilottia varten. [PAS_Pilotit_2015]
FIRE	Helsingin yliopisto, Seismologian instituutti	Ympäristö-tieteet, seismologia	Aineisto sisältää heijastusluotauksella tehtyjä seismologisia mittauksia Suomen maaperästä. Se kerättiin laajassa kansallisessa FIRE-hankkeessa vuosina 2001-2004. [FIRE_Hanke] Analysoitavaksi valittiin kaksi otosta eri luotauslinjojen mittaustuloksista, sekä niihin liittyvät oheistiedostot.

Lyhenne	Luoja / omistaja	Tieteenala	Aineiston kuvaus ja analysoitavaksi valittu osa
FSD	Yhteiskunta-tieteellinen tietoarkisto FSD	Yhteiskunta- ja humanistiset tieteet	Yhteiskuntatieteellinen tietoarkisto FSD:n keräämät ja alan tutkijoiden käyttöön tarjoamat aineistot. Osa aineistoista on avoimia, osaan tarvitaan käyttölupa. Analysoitavaksi valittiin kaksi vapaasti saatavilla olevaa kyselyaineistoa: kvalitatiivinen aineisto suomalaisten suhteesta kulttuuriperintöönsä [FSD2981] ja kvantitatiivinen aineisto suomalaisten Internetin ja median käytöstä [FSD2985].
Kiteet	Aalto-yliopisto, Biotekniikan ja kemian tekniikan laitos, Biohybridi-materiaalien tutkimusryhmä	Luonnontieteet ja tekniikka, Biokemia	Aineisto sisältää mittaustuloksia liittyen kiteiden muodostumiseen pehmeissä aineissa. Se on osa Aalto-yliopiston biotalousinfrastruktuuriin kuuluvaa materiaalitutkimusta. Analysoitavaksi valittiin tutkimusryhmän kokoama otos mittaustuloksista.
MAXIV	MAX IV -laboratorio, Lundin yliopisto, Ruotsi / Diamond Light Source Ltd, Englanti Suomessa kansallinen koordinaattori Oulun yliopisto.	Luonnontieteet ja tekniikka, Materiaali-fysiikka, lisäksi myös Bio- ja lääketieteet	Ruotsalainen MAX IV -laboratorio tarjoaa röntgenmikroskopian palveluja, joiden avulla voidaan tutkia mm. proteiinien rakennetta. [MAXIV_Lab] Analysoitavaksi valittiin laboratorion suosittelema esimerkkitiedosto Nexus HDF5 -muodossa, jossa suurin osa tiedostoista tullaan tallentamaan. Esimerkkitiedoston on luonut englantilainen synkrotronikeskus Diamond Light Source.
Planck	Euroopan avaruusjärjestö ESA	Luonnontieteet ja tekniikka, Avaruus-tutkimus	Aineisto koostuu ESAn Planck-avaruustutkimusaseman tekemistä kosmisen taustasäteilyn mittauksista reilun neljän vuoden ajalta. Se on vapaasti ladattavissa verkosta ESAn ylläpitämästä Planck Legacy Archivesta [PLA]. Analysoitavaksi valittiin yhden taajuusalueen mittausdata yhden päivän ajalta.

Lyhenne	Luoja / omistaja	Tieteenala	Aineiston kuvaus ja analysoitavaksi valittu osa
RITU	Jyväskylän yliopisto, Fysiikan laitos, Kiihdytin-laboratorio	Luonnontieteet ja tekniikka, Hiukkasfysiikka	Aineisto koostuu kiihdytinlaboratoriossa kehitetyn RITU-rekyyliseparaattorin tuottamasta mittausdatasta sekä sen oheistiedostoista. [RITU] Analysoitavana oli otos mittaustuloksista, jonka tutkijat olivat jo aiemmin valinneet vuonna 2015 tehtyä pitkäaikaissäilytyksen pilottia varten. [PAS_Pilotit_2015]
SMEAR	Helsingin yliopisto, Fysiikan laitos, SMEAR-tutkimusasemat	Ympäristö-tieteet, ilmakehätieteet	Aineisto koostuu useiden eri havaintoasemien ja mittalaitteiden tuottamasta mittausdatasta. Se sisältää mittauksia ilmakehästä, maaperästä, puustosta ja veden laadusta. Aineisto karttuu jatkuvasti ja se on tallennettu MySQL-tietokantaan. Aineisto on vapaasti tarkasteltavissa ja ladattavissa AVAA-palvelussa www-käyttöliittymän kautta. [SMEAR_AVAA] Analysoitavana oli koko tietokanta, sisältäen aineiston usealta vuodelta päivään 25.1.2016 asti.
Suomi24	CSC ja Kotimaisten kielten keskus, Kielipankki	Yhteiskunta- ja humanistiset tieteet, Kielitiede	Aineisto koostuu Suomi24-keskustelufoorumin viesteistä vuosilta 2001-2015, kielitieteellisesti jäsennettynä ja annotoituna. [Suomi24] Analysoitavaksi valittiin kolme otosta eri aikajaksoilta, yhteensä 1,5 miljoonaa viestiä koottuna yhteensä 141 tiedostoon.

5.1 Aineistojen tiedostomuodot ja koko

Aineistojen tiedostomuodot ja koko on esitetty tiivistetysti alla olevassa taulukossa.

Aineisto	Tiedostomuodot	Analysoidun osan koko Gt	Aineiston koko yhteensä
1000Gen	BAM, CRAM	34,8	useita satoja teratavuja
Aivokuvat	BIDS, JSON, NIFTI, PDF, TSV, WMV	1,3	noin 8 gigatavua
ERNE	PDF, PNG, TXT (rakenteinen)	0,8	noin 22 gigatavua
FIRE	Corel Draw, DOC, JPG, SEG-Y, TXT (rakenteinen), WMV	2,1	noin 2 teratavua

Aineisto	Tiedostomuodot	Analysoidun osan koko Gt	Aineiston koko yhteensä
FSD	PDF, RTF, SPSS Portable, TXT, XML	< 0,1	Vaihtelee, yksittäisen tutkimusprojektin aineisto tyypillisesti alle 0,1 Gt.
Kiteet	PDF, XLSX	< 0,1	Tulostiedostot alle 0,1 Gt
MAXIV	Nexus HDF5, HTML, PDF	< 0,1	Vaihtelee, sekä pieniä että suuria aineistoja (riippuu palvelua käyttävästä tutkimusprojektista).
Planck	FITS, HTML, PDF	1,1	noin 20 teratavua
RITU	Java, GREAT, PDF, TXT (rakenteinen), XML	4,6	noin 200 gigatavua
SMEAR	MySQL-tietokanta, SIARD, CSV, HDF5, HTML, JSON, TSV,	32,7	Tietokannan koko MySQL-dump-tiedostona 32,7 Gt, tietokantana reilut 10 Gt. Mittalaitteiden raakadatat mukaan lukien dataa on useita kymmeniä teratavuja.
Suomi24	TXT, VRT	4,0	Noin 170 gigatavua

Lähes jokainen aineisto koostuu useammassa eri tiedostomuodossa olevista tiedostoista. Yleisesti tunnettuja ja useilla aloilla käytettyjä ovat mm. TXT (teksti)- ja PDF (Portable Document Format) -muotoiset dokumentit, PNG (Portable Network Graphics) -muotoiset kuvat sekä WMV (Windows Media Video) -muotoiset videot. Nämä muodot on jo huomioitu Kansallisen digitaalisen kirjaston säilytys- ja siirtokelpoisten tiedostojen määrityksissä [KDK_Tiedostomuodot]. Aineistoon liittyvät julkaisut sekä muut dokumentit, kuten mittalaitteiden kuvaukset, ovat poikkeuksetta jo valmiiksi jossain KDK:ssa hyväksytyistä muodoista tai helposti niihin muunnettavissa.

Kahdeksassa esimerkkiaineistossa pääosassa on tutkimuksessa käytetyn mittalaitteen tuottama data, joko suoraan laitteen tuottamana tai vakioidun käytännön mukaisesti muokattuna. Kaksi aineistoista sisältää mittausdatan sijaan ihmisten tuottamaa tilastollista dataa (FSD, Suomi24). Yhdessä aineistossa (Kiteet) keskeiset mittaustulokset ja parametrit on koottu Excel-taulukkoon (XLSX-tiedostomuoto), jättäen mittalaitteiden tuottama raakadata pois aineistokokonaisuudesta. Datatiedostot ovat joko rakenteisia tekstitiedostoja (ERNE, VRT), binääritiedostoja (BAM, CRAM, GREAT, NIFTI, SEG-Y, SPSS Portable) tai rakenteisen tekstin ja binääridatan yhdistelmiä (FITS, Nexus HDF5). Mittauksessa käytetyt parametrit ovat useimmiten omassa tiedostossaan joko rakenteisena tekstinä, avain-arvo-pareja sisältävänä JSON-tiedostona tai XML-muodossa.

Aivokuvat-aineistossa huomionarvoista on itse tiedostojen lisäksi Brain Imaging Data Structure (BIDS) -hakemistorakenne. Aivokuvia käsittelevien tutkijoiden kansainvälisen tiedeyhteisön luoma BIDS-rakenne määrittelee käytettävien tiedostomuotojen lisäksi myös sen, miten tiedostot tulee nimetä ja minkä nimisiin hakemistoihin ne tulee tallentaa. Tarjolla on myös validointityökalu, jonka avulla BIDS-yhteensopivuus voidaan tarkastaa.

SMEAR-aineisto poikkeaa muista siten, että mittalaitteiden tuottama data on kerätty tietokantaan. Tietokannan rakenne ja sisältö voidaan tallentaa ns. dump-tiedostoon, mutta tietokannasta voidaan hakea osia aineistosta huomattavasti nopeammin ja joustavammin kuin tiedostoista. AVAA-palvelun web-hakuliittymän kautta SMEAR-aineistosta voidaan valita

haluttuja osia ja ladata ne CSV, HDF5-, tai TSV-muodossa [SMEAR_AVAA]. Tarjolla on myös JSON-rajapinta. Testausta varten aineisto muunnettiin lisäksi tietokantojen säilytystä varten kehitettyyn SIARD-muotoon.

5.2 Aineistojen metatiedot

Jotta tutkimusaineiston uudelleenkäyttö on mahdollista, tarvitaan erilaisia metatietoja. Niitä ovat mm. mittalaitteiden asetukset, koetilanteen kuvaus sekä tiedostojen rakenteen kuvaavat tiedot. Toimitettaessa aineistoja pitkäaikaissäilytykseen on varmistettava, että aineistokokonaisuus sisältää kaikki ymmärrettävyyden kannalta olennaiset metatiedot. Ymmärrettävyyden arvioinnissa on huomioitava, että esimerkillisestikään kuvattu aineisto ei usein ole maallikon ymmärrettävissä, vaan tulkinta vaatii syvällistä kyseisen tieteenalan asiantuntemusta.

Alla olevassa taulukossa on esitetty katsaus esimerkkiaineistoihin kuuluvista metatiedoista, sekä siitä miten ne on kussakin aineistossa esitetty.

Aineisto	Metatiedot
1000Gen	▪ Ymmärrettävyyden kannalta olennaista mm. geeninäytteiden prosessoinnin kuvaus, koehenkilön fenotyyppi (eräänlainen "potilasrekisteriote") ja tietyt sekvensointiin liittyvät tekniset yksityiskohdat. ▪ Mittausdatan (geenisekvenssi) sisältävissä BAM- ja CRAM-tiedostoissa on header-osa, johon metatietoja voidaan tallentaa. Osa kentistä on pakollisia, osa valinnaisia. ▪ Kansainvälisesti käytetty Sequence Read Archive (SRA) -metadatamalli [ENA_SRA] on vakioitu käytäntö geeninäytteeseen liittyvien metatietojen kuvaamiseen. ▪ Fenotyyppi tallennettu erikseen, ei vakiintunutta käytäntöä. Fenotyyppiin liittyviä tietoja ei usein voida julkaista tietosuojasyistä. Esimerkkiaineistossa fenotyyppitietoja olivat vain kansallisuus ja sukupuoli.
Aivokuvat	▪ Ymmärrettävyyden kannalta olennaista mm. aivokuvauslaitteen asetukset, koehenkilöiden fenotyyppitiedot ja tutkimusjärjestelyn kuvaus. ▪ Laitteiden asetukset ja muut olennaiset tekniset tiedot on tallennettu JSON-muotoisiin tiedostoihin, joihin tutkijaryhmä on ne laitteiden tuottamista DICOM-kuvatiedostoista siirtänyt tulkinnan ja käsittelyn helpottamiseksi. ▪ Fenotyyppitiedot on tallennettu TSV-muotoiseen tiedostoon, ei vakiintunutta käytäntöä. Kuten 1000Gen-aineistossa, fenotyyppiin liittyviä tietoja ei usein voida julkaista tietosuojasyistä. ▪ Vakioitu BIDS-hakemistorakenne ja tiedostojen nimeämiskäytäntö auttaa tutkijoita löytämään olennaiset tiedot sekä helpottaa tietojen automaattista käsittelyä. ▪ Lyhyt aineiston kuvaus sekä muita vakioituja tietokenttiä JSON-muotoisessa dataset_description.json-tiedostossa (osa BIDS-määritystä). ▪ Tutkimuksen tarkempi kuvaus PDF-muotoisissa artikkeleissa.

Aineisto	Metatiedot
ERNE	▪ Ymmärrettävyyden kannalta olennaista mm. mittalaitteen toiminnan kuvaus, laitteen asetukset ja koetilanne. ▪ Datatiedostoissa itsessään on header-osa, jossa tiedoston rakenne on kuvattu ja sen sisältämät suureet lueteltu. ▪ Mittalaitteen asetukset, koetilanne, ym. datan tulkintaan vaikuttavat tiedot, kuten myös itse mittalaitteen kuvaus ovat erillisissä PDF-muotoisissa dokumenteissa. ▪ Datan tieteellisessä tulkinnassa käytettyjä muita aineistoja (esim. aurinkotuulen magneettikenttä, satelliitin tarkka paikka ja asento) ei ole liitetty aineistoon.
FIRE	▪ Ymmärrettävyyden kannalta olennaista mm. mittalaitteiden asetukset, havaintopisteiden koordinaatit, yksittäisten mittausten tiedot sisältävä havaintopäiväkirja ja kenttätyöraportti (kuratoijan raportti), joka sisältää mittauksen kuvailun, parametrit ym. ▪ Mittausdatan sisältävissä SEG-Y-tiedostoissa on header-osa, johon tallennetaan tietyt vakioidut mittaukseen liittyvät metatiedot. ▪ Havaintopisteiden koordinaatit sekä havaintopäiväkirja ovat rakenteisissa tekstitiedostoissa, ei vakiintunutta käytäntöä tallentamisen yksityiskohdista. ▪ Kenttätyöraportti DOC-muotoinen dokumentti ▪ Mittaushanke ja siitä saatuja tuloksia on esitelty sekä kirjallisesti PDF-muotoisissa artikkeleissa sekä WMV-muotoon tallennetussa videossa.
FSD	▪ Ymmärrettävyyden kannalta olennaista mm. aineiston sisällön kuvaus, tiedot kyselyyn vastanneesta ja/tai tutkimuksen kohteena olleesta henkilöryhmästä sekä analyysissa käytetyt muuttujat. ▪ Sisällön kuvaus, lista muuttujista ja muut olennaisimmat metatiedot ovat koneluettavassa DDI 2.0 -standardin mukaisessa XML-muodossa, ja lisäksi helpommin ihmisen luettavissa olevassa PDF-dokumentissa. ▪ Esimerkkiaineisto on FSD:n jo valmiiksi säilytystä ja uudelleenkäyttöä varten työstämä kokonaisuus. Yhteiskuntatieteiden tutkijat eivät tyypillisesti itse tuota metatietoja yhtä organisoidussa muodossa, vaan alalla on yleinen käytäntö, että aineisto käsitellään ja metatiedot yhtenäistetään tietoarkiston tarjoamana palveluna.
Kiteet	▪ Ymmärrettävyyden kannalta olennaista tutkimusmenetelmä, joka on kuvattu aineiston osana PDF-muodossa toimitetussa tieteellisessä julkaisussa. ▪ Aineistossa ei ole mukana mittalaitteiden tuottamaa raakadataa. Keskeiset mittaustulokset sekä käytetyt parametrit on koottu yhteen Excel-tiedostoon. Osa tuloksista on esitetty numeroarvojen lisäksi myös graafisina kaavioina. ▪ Aineisto on tarkoitettu lähinnä ihmissilmin tarkasteltavaksi. Valitsemalla haluttuja osia tuloksista ja tallentamalla ne erillisinä taulukoina voidaan tuottaa koneluettavia tiedostoja.

Aineisto	Metatiedot
MAXIV	▪ Esimerkkiaineisto ei ole minkään varsinaisen tutkimusprojektin aineisto, koska MAX IV -laboratorion toiminta ei ole vielä käynnistynyt. ▪ Metatiedot on tallennettu Nexus HDF5 määrityksen mukaisesti binäärimuodossa HDF5-muotoiseen tiedostoon.
Planck	▪ Ymmärrettävyyden kannalta olennaista mm. mittalaitteen (Planck-satelliitti) kuvaus, asetukset sekä menetelmä miten aineisto on tuotettu ▪ Mittausdatan sisältävissä FITS-tiedostoissa on header-osa, johon on tallennettu mittauksessa käytetyt asetukset ja parametrit. ▪ Mittalaitteen sekä aineiston tuottamiseen käytetyn menetelmän kuvaukset ovat HTML-muodossa datatiedostot sisältävän arkiston verkkosivuilla.
RITU	▪ Ymmärrettävyyden kannalta olennaista mm. mittalaitteen kuvaus, konfiguraatioparametrit, havaintopäiväkirja ja datatiedoston rakenteen kuvaus ▪ Mittalaitteen kuvaus on tekstinä tekstitiedostossa sekä PDF-muotoon tallennettuna kaaviokuvana. ▪ Konfiguraatioparametrit on tallennettu rakenteiseen tekstitiedostoon. ▪ Datatiedoston rakenteen kuvaus on PDF-muotoinen dokumentti. ▪ Mittalaitteen GREAT-tiedostomuoto on valmistajan oma, alalla ei vakiintunutta käytäntöä mittausdatan tiedostomuodoista. ▪ Analyysin tulokset on tallennettu Aida XML-muodossa, joka on alalla yleisesti käytössä oleva tiedostomuoto.
SMEAR	▪ Ymmärrettävyyden kannalta olennaista mm. käytettyjen mittalaitteiden kuvaukset, havaintoasemien sijainnit, mittausparametrit ja tietokannan rakenne ▪ Mittalaitteiden kuvaukset ovat HTML-muodossa SMEAR-projektin verkkosivuilla. ▪ Havaintoasemien sijainnit, mitattavien suureiden lyhyet kuvaukset sekä maininnat datalle tehdyistä toimenpiteistä (esim. jälkikäsittely, laatutarkistus) ovat tietokannassa erillisissä tauluissa.
Suomi24	▪ Ymmärrettävyyden kannalta olennaista mm. lähdeaineiston taustatiedot sekä datan (sanojen ja lauseiden) jäsentelyssä käytettyjen lyhenteiden merkitys ▪ Lähdeaineiston suppeat perustiedot ovat tekstitiedostossa. ▪ Varsinaisen datan sisältävissä VRT-tiedostoissa on kunkin keskustelualueen viestiin liittyvät metatiedot XML:ää muistuttavassa rakenteessa. ▪ Datan jäsentelyssä käytettyjä lyhenteitä ei ole dokumentoitu. Kielitieteen asiantuntija pystyy haastatellun tutkijan mukaan päättelemään mitä ne tarkoittavat.

Kansallisen digitaalisen kirjaston standardisalkussa [KDK_Standardisalkku] metatiedot jaotellaan kuvaileviin, hallinnollisiin ja rakenteellisiin metatietoihin. Hallinnolliset metatiedot jakautuvat edelleen teknisiin metatietoihin, pitkäaikaissäilytyksen metatietoihin ja käyttöoikeustietoihin.

Standardisalkussa on lueteltu suositeltavat kuvailevan metatiedon formaatit. Tutkimusaineistojen osalta vastaavia, laajasti käytettyjä formaatteja on löydettävissä vain joidenkin aineistojen ja tutkimusalojen osalta. Esimerkiksi yhteiskuntatieteiden aineistoissa DDI-formaatti on alalla yleinen ja laajan kansainvälisen yhteenliittymän standardoima, joten se voidaan valita suositeltavaksi formaatiksi myös Tutkimus-PASissa. Monien muiden alojen kohdalla joudutaan selvittämään, löytyykö aineistojen kuvailuun soveltuvia yleiskäyttöisiä formaatteja ja millaiset kriteerit kuvailulle tulee asettaa.

Tekniset metatiedot liittyvät kiinteästi tiedostomuotoihin. KDK:ssa hyväksytyt teknisten metatietojen metatietoskeemat on lueteltu Säilytys- ja siirtokelpoiset tiedostomuodot -dokumentissa [KDK_Tiedostomuodot]. Tutkimusaineistojen osalta vastaavia, valmiita metatietoskeemoja on olemassa vain harvoille tiedostomuodoille. Joiltain osin voidaan tehdä yleisiä kaikki aineistot kattavia linjauksia, kuten mitä merkistöä teksteissä tulisi käyttää, mutta monet metatiedoista liittyvät vain johonkin tiettyyn tiedostomuotoon, tieteenalaan tai tutkimusmenetelmään.

Pitkäaikaissäilytyksen metatiedot, käyttöoikeustiedot sekä rakenteelliset metatiedot voitaneen myös tutkimusaineistojen osalta tallentaa KDK:ssa jo määritellyissä formaateissa (PREMIS ja METS). Asiaan ei kuitenkaan ole tässä selvityksessä perehdytty tarkemmin. Tarkempaa tietoa näiden formaattien soveltuvuudesta ja mahdollisista muutostarpeista saadaan kuvailu- ja paketointipalvelun suunnittelun yhteydessä sekä pilotoitaessa aineistojen paketointia pitkäaikaissäilytystä varten. Vuonna 2015 toteutetuissa piloteissa havaittiin muun muassa, että omistus- ja käyttöoikeustietojen merkitsemiseen tulee kiinnittää erityistä huomiota [PAS_Pilotit_2015]. Aineisto voi esimerkiksi itse datan osalta olla vapaasti käytettävissä, mutta ymmärrettävyyden säilymisen kannalta olennaiset, tutkimusmenetelmän kuvaavat julkaisut kustantajan tekijänoikeuden alaisia.

Aineistojen metatietoihin ja paketointiin liittyviä vaatimuksia käsitellään luvussa Aineistojen hyväksyminen pitkäaikaissäilytykseen. Käyttöoikeuksiin ja niiden hallintaan liittyviin metatietoihin on perehdytty tarkemmin erillisessä, lähiaikoina julkaistavassa selvityksessä.

6 TIEDOSTOMUOTOJEN ANALYYSI

6.1 Esimerkkiaineistoissa esiintyvät tiedostomuodot

Esimerkkiaineistoissa esiintyi yhteensä 26 eri tiedostomuotoa, joista puolet on hyväksytty KDK-PAS:ssa säilytys- tai siirtokelpoisiksi muodoiksi. Säilytyskelpoisia muotoja oli yhteensä 10 kpl: HTML, Java (säilytettävissä tekstinä), JPEG, JSON (säilytettävissä tekstinä), PDF, PNG, TSV (säilytettävissä tekstinä), TXT (normaali ja rakenteinen tekstitiedosto) ja XML. Siirtokelpoisia muotoja olivat DOC/DOCX, WMV ja XLSX eli yhteensä kolme eri tiedostomuotoa.

Lopuista, KDK-PAS:ssa hyväksymättömistä tiedostomuodoista valtaosa eli 11 kpl oli avoimia ja dokumentoituja: BAM/SAM, CRAM, FITS, GREAT, HDF5, MySQL dump, NIfTI, RTF, SEG-Y, SIARD ja VRT. Suljettuja muotoja oli kaksi, CorelDraw- sekä SPSS Portable -muodot. Lisäksi Aivokuvat-aineisto oli järjestetty alalla käytetyn BIDS-määrityksen mukaisesti, joka ei ole varsinainen tiedostomuoto vaan hakemistorakenne.

Suurin osa tämän otoksen tiedostomuodoista on joko jo valmiiksi hyväksyttävissä muodoissa tai ne voitaisiin lisätä hyväksyttävien listaan määrittelemällä vaadittavat tekniset metatiedot ja muut yksityiskohdat. Toisaalta on huomioitava, että osa muodoista (Java, JSON, TSV) on KDK-PAS:ssa tuettu vain normaalina tekstinä — tukea olisi mahdollista parantaa. Myös itse kehitetyt tai tutkimusalakohtaiset rakenteiset tekstitiedostot ovat sinänsä säilytettävissä tekstinä, mutta niiden oheen tulee liittää rakenteen dokumentointi.

Suljetuista tiedostomuodoista FIRE-aineistossa esiintyvät CorelDraw-tiedostot olisi pienellä vaivalla mahdollista muuntaa PDF-muotoon. Ne ovat aineistoon liittyvää dokumentaatiota, jota käyttäjän tulee päästä lukemaan, mutta muokattavuus ei ole olennaista. FSD-aineiston SPSS Portable -muotoisia tiedostoja puolestaan on tarve jatkokäytössä päästä myös muokkaamaan, eikä niiden muuntaminen häviöttömästi avoimeen muotoon ole suoraviivaista.

Vertailussa kuuden muun aineistoja säilyttävän organisaation laatimiin säilytys- ja siirtokelpoisten tiedostomuotojen listoihin päästiin samansuuntaiseen tulokseen. KDK-PAS:ssa hyväksytyt muodot olivat myös laajasti kansainvälisesti hyväksyttyjä. KDK:ssa toistaiseksi hyväksymättömistä tiedostomuodoista neljä (HDF5, RTF, SIARD ja SPSS Portable) oli osassa organisaatioista hyväksyttyjen joukossa, loput yhdeksän puuttuivat myös kaikilta kansainvälisiltä listoilta.

Esimerkkiaineistojen tiedostomuodot on analysoitu tarkemmin liitteessä C.

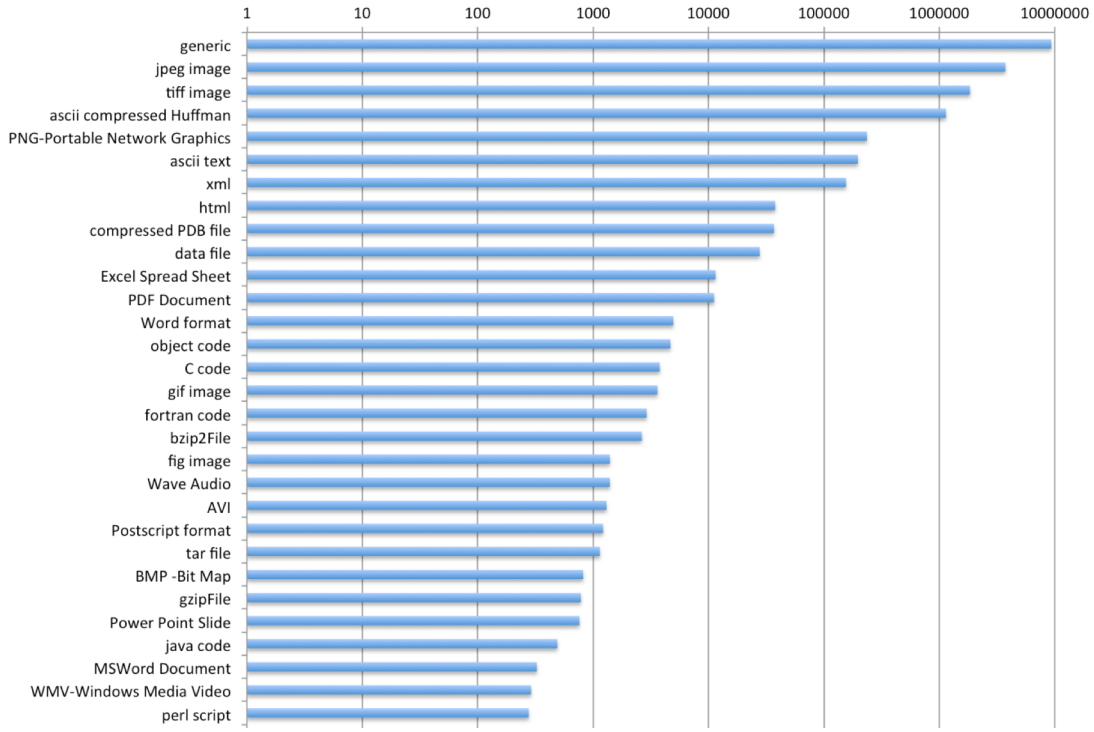
6.2 Tiedostomuotojen jakauman määrällinen kartoitus

Tiedostomuotojen yleisyyttä suomalaisessa tutkimuksessa kartoitettiin karkeasti monien tutkijoiden käyttämän CSC:n IDA-tallennuspalvelun pohjalta. Kolmekymmentä yleisintä IDAan tallennettua tiedostomuotoa on esitetty kuvassa 2.

Kuvasta nähdään, että monet KDK:ssa pitkäaikaissäilytykseen hyväksytyistä tiedostomuodoista esiintyvät usein myös IDAan tallennetuissa tutkimusaineistoissa. Esimerkiksi JPEG-, PNG- ja TIFF-kuvat, Excel-, Word ja PDF-dokumentit, ASCII-tekstit sekä XML-tiedostot on KDK:ssa hyväksytty joko siirto- tai säilytyskelpoisiksi.

Toisaalta kuvasta ilmenee suuri määrä tiedostomuotoja, joita KDK:ssa ei ole vielä huomioitu. Kaikkein yleisin tyyppi on "generic", joka tarkoittaa yksinkertaisesti että IDA-palvelun automaattinen analyysi ei tunnistanut muotoa. Monet niistä ovat todennäköisesti mittaustuloksia tai muita datatiedostoja. Myös TAR- ja GZIP-tiedostot sisältävät useita eri tiedostomuotoja, koska analyysi ei ulotu TAR- ja GZIP-pakettien sisälle. Tunnistetuista muodoista usein esiintyviä ovat mm. ohjelmien lähdekooditiedostot (C, Fortran, Java ja Perl).

Kattavampi analyysi tutkimuksen tiedostomuotojen yleisyydestä vaatisi laajaa yliopistoille ja tutkimusryhmille suunnattua kyselyä, jollaista ei tämän selvityksen puitteissa ollut mahdollista toteuttaa. Itävallassa on tehty laaja, koko tutkimuskentän kattava selvitys [Austrian_Survey], jonka tulokset ovat saman suuntaisia IDAan pohjautuvan katsauksen



Kuva 2: 30 yleisintä IDA-palvelussa esiintyvää tiedostomuotoa

kanssa. Lähes kaikki tutkijat tuottavat tekstiä, taulukoita sekä kuvia, mutta myös itse kehitetyt ohjelmat (lähdekoodi sekä suoritettavat tiedostot) ja mittausdata ovat merkittäviä tiedostoryhmiä. Lisäksi Itävallan selvityksessä esille nousevat tietokannat, joiden osalta jää kuitenkin epäselväksi, millaisista tietokannoista on kyse. Ainakin osa niistä lienee yhteiskuntatieteissä yleisesti käytettyjen tilasto-ohjelmien tuottamia tiedostoja, joille ei kyselyssä ollut muuta sopivaa luokkaa.

Määrään pohjautuvassa analyysissa korostuvat tiedostomuodot, joissa data on jaettu useaan pieneen tiedostoon yhden suuremman sijaan. Toisaalta tiedostojen kokoon tukeutuminen korostaisi suuria datamassoja käsitteleviä tieteenaloja. Säilytyksen ja uudelleenkäytön kannalta kooltaan pienempiä aineistoja tuottavat alat ovat kuitenkin yhtä tärkeitä. Itävallan selvitys antaa hiukan enemmän tietoa siitä, mitä tiedostomuotoja tukemalla voidaan palvella mahdollisimman monia tutkijoita. Sitäkin voidaan kuitenkin hyödyntää vain yleisellä tasolla, haluttaessa tietää tarpeista tarkemmin on perehdyttävä tiedostomuotoihin tieteenala-kohtaisesti.

6.3 Tutkimusaineistoissa usein käytettyjä tiedostomuotoja

Tässä osiossa esitellään tutkimusaineistoissa usein käytettyjä tiedostomuotoja käyttötarkoituksen ja tieteenalojen mukaan jaoteltuina. Tiedot perustuvat projektissa tehtyihin haastatteluihin ja eri alojen verkkosivuihin. Otos ei kata kaikkia aineistotyyppejä eikä tieteenaloja, mutta antaa hyvän yleiskuvan erilaisista aineistoista ja niiden tiedostomuodoista pitkäaikaissäilytyksen ja -saatavuuden näkökulmasta.

Yleiskäyttöiset tieteellisen datan tiedostomuodot

Yleiskäyttöiset tieteellisen datan tiedostomuodot tarjoavat mahdollisuuden tallentaa mm. liukulukuja sisältäviä taulukoita ja muita tieteellisessä datassa usein esiintyviä rakenteita tehokkaasti ja laitealustariippumattomasti, jolloin tiedostot ovat mm. eri tavujärjestystä käyttävien tietokonearkkitehtuurien (big endian vs. little endian) välillä yhteensopivia. Osan muodoista kehitys on saanut alkunsa jollakin tietyllä tieteenalalla, mutta itse tiedostomuodon rakenne ja määritys on tieteenalasta riippumaton ja yleiskäyttöinen.

Tunnetuin tähän ryhmään kuuluva tiedostomuoto on Hierarchical Data Format 5 (HDF5). Se määrittelee kaksi peruselementtiä, joita hyödyntämällä voidaan tallentaa lähes minkä tahansa tyyppistä dataa ja niihin liittyviä metatietoja, sekä järjestää dataobjektit halutulla tavalla

puumaiseen rakenteeseen. HDF5 on avoin standardi, mutta määritys on varsin pitkä ja monimutkainen. Pitkäaikaissäilytyksen kannalta on lisäksi huomioitava, että HDF5-muodon käyttö ei sellaisenaan takaa ymmärrettävyyttä, vaan datatyypit ja metatiedot on myös määriteltävä.

HDF5:n päälle on luotu eri hankkeissa tarkentavia määrityksiä, joissa kuvataan tietyt, esimerkiksi tietyllä tieteenalalla tai tietynkaltaisissa aineistoissa tarvittavat datatyypit. Tällaisia HDF5:een pohjautuvia tiedostomuotoja ovat mm. Network Common Data Form versio 4 (NetCDF-4), Data Exchange [DXFile] ja MAX IV -laboratorion esimerkkiaineistossa sekä useissa muissa synkrotronikeskuksissa käytetty Nexus HDF5. Ne ovat pitkäaikaissäilytyksen kannalta helpommin hallittavia kuin täysin yleiskäyttöinen HDF5, johtuen siitä että sallitut datatyypit on määritelty tarkemmin. Toisaalta HDF5:n variaatioiden käsittely erillisinä tiedostomuotoina johtaa suurempaan lukumäärään muotoja, niiden määrityksiä ja versioita.

Vanhempia, mutta edelleen laajasti käytettyjä sekä ylläpidettyjä, yleiskäyttöisiä tieteellisen datan tiedostomuotoja ovat Common Data Format (CDF), Network Common Data Form versio 3 (NetCDF-3) ja Hierarchical Data Format 4 (HDF4). Niitä voidaan käyttää samantyyppisten aineistojen tallentamiseen, mutta ominaisuuksissa ja rakenteissa on osin merkittäviä eroja eivätkä muodot siksi ole keskenään yhteensopivia — eivät edes NetCDF-3 ja NetCDF-4 tai HDF4 ja HDF5 [CDF_FAQ].

Pitkäaikaissäilytyksen kannalta yleiskäyttöiset tiedostomuodot vastaavat sikäli tutkijoiden itse määrittelemiä muotoja, ettei ymmärrettävyyden säilytyksen ja uudelleenkäytön kannalta olennaisille metatiedoille ole tiedostomuotojen määrityksissä yksikäsitteistä paikkaa. Toisaalta standardointi edes yleisellä tasolla ja tiedostojen käsittelyä helpottavat, vapaasti saatavilla olevat ohjelmointikirjastot ovat uudelleenkäytön kannalta merkittävä etu kokonaan itse määriteltyihin muotoihin verrattuna. Siksi yleiskäyttöisten tiedostomuotojen käyttöä kannattaa myös PAS-ratkaisussa suosia sekä laatia kriteerit ja työkalut, joiden avulla käytetyt rakenteet ja metatiedot voidaan dokumentoida ymmärrettävyyden säilytyksen kannalta riittävällä tarkkuudella.

Kansainvälisesti HDF5 on hyväksytty siirto- tai säilytyskelpoiseksi kolmessa (CINES, DANS, LoC) kuudesta tarkastellusta organisaatiosta. LoC mainitsee lisäksi CDF:n. Tuen tasosta tai muotoihin liittyvistä dokumentointivaatimuksista ei ollut tarkempaa tietoa.

Mittalaitekohtaiset tiedostomuodot

Mittalaitekohtaisia tiedostomuotoja käytetään laajasti useilla eri tutkimusalueilla. Ne poikkeavat toisistaan hyvin paljon, johtuen sekä itse laitteiden erilaisuudesta että laitevalmistajien vaihtelevista käytännöistä. Pääasiassa yrityksille tuotantokäyttöön myytävien mittalaitteiden tiedostomuodot ovat usein suljettuja ja niiden lukemiseen tarvitaan erikoisohjelmisto. Alun perin tutkimusta varten kehitettyjen mittalaitteiden tuottamat tiedostot ovat usein niin ikään laite- tai alakohtaisia, mutta avoimesti dokumentoituja.

Mittauksessa käytetyt parametrit sekä laitekohtaiset metatiedot ovat itse datan lisäksi olennaisia aineiston tulkinnan kannalta. Ne saatetaan tallentaa joko samaan tiedostoon mittausdatan kanssa (tiedoston header-osaan) tai erillisiin tiedostoihin. Pitkäaikaissäilytyksen kannalta on tärkeää tunnistaa, mitkä parametrit ja metatiedot ovat ymmärrettävyyden ja uudelleenkäytön kannalta olennaisia, sekä huolehtia että vastaanotettu aineisto sisältää kyseiset tiedot. Myös mittalaitteen toimintaperiaatteen kuvaus saattaa olla tarpeen aineiston ymmärtämiseksi.

Esimerkkejä mittalaitekohtaisista tiedostomuodoista ovat RITU-esimerkkiaineistossa käytetty GREAT-muoto sekä elektronimikroskooppien tuottamat Digital Micrograph 3 (DM3) -tiedostot. Ensiksi mainitun dokumentaatio on ladattavissa valmistajan sivuilta, DM3-tiedostojen tapauksessa käyttäjät ovat itse tutkineet tiedostojen rakennetta ja laatineet havaintojensa pohjalta muodon osittain kuvaavan dokumentaation.

Mittausdatan tyypistä riippuen saattaa olla mahdollista muuntaa tiedostot helpommin säilytettävään muotoon. Yleisesti käytetyt muodot helpottavat myös aineistojen

hyödyntämistä tieteenalarajojen yli. Esimerkiksi elektronimikroskooppien tiedostot ovat pohjimmiltaan bittikarttakuvia ja siten tallennettavissa esimerkiksi KDK:ssa säilytyskelpoiseksi hyväksytyssä TIFF-muodossa. Tällöin kuitenkin menetetään DM3-tiedoston sisältämät metatiedot, jotka on näin meneteltäessä tallennettava erikseen. Muunnoksessa on lisäksi kiinnitettävä huomiota siihen, että kuvan alkuperäinen tarkkuus ja bittisyvyys säilyvät.

Projektissa tarkasteltujen ulkomaisten säilytysorganisaatioiden tiedostomuotojen listoilla mittalaitekohtaisten tiedostomuotojen säilytykseen ei oteta kantaa tai tarjota aineeseen liittyvää ohjeistusta.

Paikkatietoaineistojen tiedostomuodot

Paikkatietoaineistot ja kartat ovat pitkäaikaissäilytyksen osalta erityisen kiinnostavia, koska niitä voidaan hyödyntää useilla eri tieteenaloilla sekä tieteenalarajat ylittävässä tutkimuksessa. Maantieteelliset koordinaatit sisältäviä mittaustuloksia ja muuta dataa kuten valtioihin tai kuntiin liittyviä tilastoja voidaan vertailla keskenään sekä sijoittaa erilaisille karttapohjille visuaalista tarkastelua varten. Olennaisten piirteiden kuten koordinaattijärjestelmien yhteensopivuus on aineistoja vertailtaessa erityisen tärkeää.

Alalla käytetyt tiedostomuodot voidaan karkeasti jakaa vektoripohjaisiin ja rasteripohjaisiin muotoihin. Vektoripohjaiset muodot pohjautuvat koordinaatteihin sekä niitä yhdistäviin suoriin tai kaareviin viivoihin, rasteripohjaiset muodot puolestaan tasaväliseen hilaan, jossa pisteet ovat kiinteällä etäisyydellä toisistaan. Molempiin tarkoituksiin on tarjolla useita eri tiedostomuotoja. Vektoripohjaisia muotoja ovat mm. Esri Shapefile (Shape), Geography Markup Language (GML) ja Keyhole Markup Language (KML). Rasteripohjaisia muotoja ovat mm. GeoTIFF, JPEG2000 ja PNG. Lisäksi on vielä joitakin näihin ryhmiin kuulumattomia muotoja, kuten laserkeilauksessa käytetty LAS ja erilaiset tietokannat.

Esri Shapefile on yksityisen paikkatieto-ohjelmistoja myyvän yrityksen kehittämä joukko toisiinsa liittyviä tiedostomuotoja, joista on ohjelmistojen suosion (n. 40% osuus paikkatieto-ohjelmistojen markkinoista) ansiosta muodostunut alalla de facto -standardi. Muoto on yksinkertainen, stabiili, melko hyvin dokumentoitu ja tuettu myös muissa kuin Esrin kehittämissä ohjelmistoissa, soveltuen siten pitkäaikaissäilytyksen tarpeisiin. Shapefile-muoto on jo käytössä mm. Avaa-portaalista löytyvän Paituli-paikkatietopalvelun aineistoissa. Pitkäaikaissäilytyksen osalta on kuitenkin syytä tarkentaa, mitkä Shapefilen valinnaisista ominaisuuksista ovat tuettuja, ja mitkä metatiedot säilytettäviltä aineistoilta vaaditaan. Lisäksi on tarpeen huolehtia, että vastaanotettu kokonaisuus sisältää kaikki tarpeelliset osat: Shapefile-muoto koostuu useista erillisistä, toisiinsa kuuluvista tiedostoista.

Open Geospatial Consortium (OGC) on paikkatietoaineistoihin keskittyvä, vapaaehtoiseen osallistumiseen perustuva standardointijärjestö, johon on liittynyt jäseniksi yli 500 organisaatiota. Jäsenistöön kuuluu sekä kaupallisia yrityksiä että ei-kaupallisia järjestöjä, valtiollisia elimiä ja tutkimusorganisaatioita. OGC on laatinut tai valinnut suosituslistaansa useita kymmeniä toisiaan täydentäviä paikkatietoihin liittyviä standardeja. Ne ovat kaikki vapaasti saatavilla järjestön verkkosivuilta [OGC_Standards].

Tutkimusaineistojen kannalta tärkein OGC:n standardi on Geography Markup Language (GML), XML-pohjainen merkintäkieli erilaisten paikkatietoa sisältävien piirteiden esittämiseen. Se on myös ISO-standardi (ISO 19136:2007). Standardin perusosan lisäksi GML-tiedostoissa voidaan käyttää yhteisön kehittämiä laajennuksia. GML on avoin, hyvin dokumentoitu ja laajasti tuettu, soveltuen siten pitkäaikaissäilytettäväksi. GML:n laajennukset voidaan hyväksyä säilytyskelpoisina XML-dokumentteina, vaikkei niille erillistä pitkäaikaissäilytyksen tukea olisikaan. Geography Markup Language -muotoa ei tule sekoittaa vanhempaan, graafien tallentamiseen tarkoitettuun Graph Modeling Language -tiedostomuotoon, josta käytetään samaa lyhennettä GML.

Keyhole Markup Language (KML) on Googlen kehittämä XML-pohjainen merkintäkieli, joka on tarkoitettu erityisesti kaksi- ja kolmiulotteisten karttojen annotointiin ja visualisointiin. Se on nykyisin myös OGC:n hyväksymä standardi. KML on osittain päällekkäinen GML:n kanssa ja jatkossa kieliä on tarkoitus yhtenäistää tai ainakin parantaa niiden keskinäistä

yhteensopivuutta. KML on jo nykymuodossaan hyvin dokumentoitu ja soveltuu pitkäaikaissäilytettäväksi, ainakin KDK:ssa jo säilytyskelpoiseksi hyväksyttynä XML-dokumenttina ilman erityistä KML-tukea.

Kolmas OGC:n yleinen datastandardi on GeoPackage. GeoPackage voi sisältää sekä vektori- että rasteriaineistoja. Teknisesti kyseessä on SQLite-tietokanta. GeoPackage on uusi standardi ja siitä on kaavailtu Shapen ja GML/KML:n korvaajaa, mutta toistaiseksi on se vielä suhteellisen harvoin käytetty muoto.

Rasteripohjaisista muodoista tärkein on Tagged Image File Format (TIFF) -kuvatiedostomuoto, joka on KDK:ssa jo hyväksytty säilytyskelpoiseksi tiedostomuodoksi. Paikkatietoaineistojen TIFF-kuvissa voi kuitenkin olla peruskuvan lisäksi ylimääräisiä kanavia tai pieniä lisätiedostoja, johon on tallennettu esimerkiksi kuvan sijainti ja koordinaattijärjestelmä. GeoTIFF-standardi puolestaan määrittelee paikkatietoaiheisten metatietokenttien tallennuksen TIFF-kuvien yhteyteen. Sekä monikanavaisten TIFF-kuvien että GeoTIFF-metatietojen käsittely on syytä huomioida TIFF-kuvien pitkäaikaissäilytyksen tuen osalta. Muita paikkatietoaineistoissa käytettyjä rasteripohjaisia muotoja ovat KDK:ssa jo säilytyskelpoisiksi hyväksytyt JPEG2000 ja PNG.

Laserkeilaukseen on oma tiedostomuotonsa, alalla de facto -standardiksi muodostunut LASer (LAS) -muoto. Se on melko yksinkertainen binäärimuoto, joka sisältää header- ja data-osiot. Header-osio sisältää tärkeimmät mittaukseen liittyvät metatiedot. LAS-muoto on avoin, hyvin dokumentoitu sekä tuettu laajasti alan ohjelmistoissa, ja soveltuu siten pitkäaikaissäilytettäväksi.

Paikkatietoaineistoja tallennetaan yhä enenevässä määrin erilaisiin tietokantoihin, jotka tarjoavat nopean ja kätevän tavan valita haluttu osa aineistosta sekä tehokkaat hakutoiminnot. Standardia tietokantamuotoa ei ole, mikä tekee niistä pitkäaikaissäilytyksen näkökulmasta muita muotoja hankalampia. Tietokantoja on käsitelty tarkemmin osiossa Tietokantamuotoiset aineistot.

Paikkatietoaineistojen yhteensopivuuden kannalta olennaista on tiedostomuotojen lisäksi koordinaattijärjestelmien valinta. Maailmanlaajuisesti on olemassa jopa kymmeniä tuhansia eri koordinaattijärjestelmiä, ja Suomessakin voidaan käyttää jopa kuntakohtaisesti eri koordinaatistoja. Aineistojen uudelleenkäytön helpottamiseksi on syytä valita mahdollisimman pieni joukko tuettavia koordinaattijärjestelmiä, ja vaatia että vastaanotettavissa aineistoissa käytetään jotain niistä. Suomen aineistojen osalta tulisi käyttää JHS-197 -suosituksen mukaisia koordinaattijärjestelmiä, ensisijaisesti ETRS-TM35FIN-koordinaatistoa.

Kansainvälisesti paikkatietoaineistoihin on kiinnitetty ainakin jonkin verran huomiota kaikissa kuudessa tarkastellussa organisaatiossa. Valinnat ja suositukset poikkeavat melko paljon toisistaan. CINESissä ainoa hyväksytty muoto on GeoTIFF, joka löytyy myös kaikilta muilta paitsi NAA:n listalta. Open Geospatial Consortiumin GML on hyväksytty DANSissa, LAC:ssa, LoC:ssa ja UKDA:ssa. ESRI Shapefile ja KML löytyvät DANSin, LAC:n ja UKDA:n listoilta, NAA puolestaan suosittelee Autodeskin kehittämää Spatial Data File (SDF) -muotoa. LAC:n listalla on vielä melkoinen joukko lisää muotoja. LoC suosittelee tallentamaan mahdollisimman täydellisen alkuperäisen aineiston, vaikka muoto olisikin suljettu, ja suosittelee lisäksi yleisesti laajasti käytettyjen GIS-sovellusten natiivimuotoja sekä OGC:n laatimia tai valitsemia tiedostomuotoja.

Ohjelmien lähdekoodi- ja binääritiedostot

Käytännössä kaikilla tutkimusaloilla ainakin osa tutkijoista ohjelmoi itse, ja aineistoihin liittyy itse kehitettyjen ohjelmien lähdekoodi- ja binääritiedostoja. Niiden säilytys on hyödyllistä sekä tutkimuksen toistettavuuden varmistamiseksi että uudelleenkäytön kannalta: datan analyysia varten kehitetyillä ohjelmilla pääsee myös jatkotutkimuksessa usein nopeimmin alkuun.

Lähdekooditiedostot ovat pitkäaikaissäilytyksen kannalta periaatteessa helppoja, ne ovat KDK-PAS:ssa jo hyväksyttyjä tekstitiedostoja, käytetystä ohjelmointikielestä riippumatta.

Metatietoihin on kuitenkin syytä kiinnittää huomiota, esimerkiksi käytetyn ohjelmointikielen nimi ja versio ovat olennaisia tietoja. Ohjelmakoodin sisäisen dokumentaation laatu vaihtelee huomattavasti, mutta laadun arviointi on käytännössä mahdotonta: se vaatisi tiedostojen yksityiskohtaista manuaalista tarkastelua ja ohjelmaan perehtymistä. Dokumentaatio voidaan haluttaessa koneellisesti eritellä itse koodista ja indeksoida hakutoimintoja varten.

Lähdekoodista käännetyt, suoritettavat ohjelmatiedostot (binääritiedostot) ovat käyttäjille käteviä, mutta pitkäaikaissäilytyksen näkökulmasta hankalia. Niiden toimivuus riippuu tyypillisesti sekä käyttöjärjestelmästä että suuresta joukosta apukirjastoja, usein edellytyksenä on jopa tietyt versiot kyseisistä kirjastoista. Ohjelmatiedostoja voi olla hyödyllistä ottaa vastaan ja tarjota ladattavaksi aineiston ohessa, mutta toimintatakuuta tulevissa järjestelmäversioissa niille ei voi antaa.

Myös lähdekoodin kääntäminen suoritettavaksi ohjelmaksi saattaa olla hankalaa, jos käytössä on alkuperäistä kehitysympäristöä uudempi järjestelmä uusine kirjastoineen. Lähdekoodia on kuitenkin binääritiedostoista poiketen mahdollista muokata, mikä antaa osaavalle käyttäjälle mahdollisuuden tehdä tarvittavat muutokset kääntämisen onnistumiseksi. Lisäksi lähdekoodin silmämääräinen tarkastelu voi helpottaa aineiston tai tutkimusmenetelmän ymmärtämistä. Siksi ohjelmien lähdekooditiedostot kannattaa sisällyttää osaksi säilytettävää aineistokokonaisuutta.

Tarkastelluissa ulkomaisissa säilytysorganisaatioissa ohjelmien lähdekooditiedostot ovat KDK:n tapaan säilytettävissä tekstitiedostoina. Vain LoC:n määrityksissä annetaan yksityiskohtaisempia ohjeita koodiin liittyvien metatietojen sekä käyttöjärjestelmäympäristön kuvaamiseksi ja säilyttämiseksi.

Merkintäkielet

Merkintäkieliä voidaan käyttää tiedealasta riippumatta moneen eri tarkoitukseen: niiden avulla voidaan tallentaa itse dataa, metatietoja tai kirjoittaa dokumentaatiota. Suosittuja merkintäkieliä ovat mm. HTML, JSON ja erityisen monikäyttöinen XML, jotka on esitelty tarkemmin liitteessä C. Muita varteenotettavia kieliä ovat Standard Generalized Markup Language SGML, artikkelien ja kirjojen kirjoittamiseen tarkoitettu LaTeX ja erityisesti metatietojen esittämiseen soveltuva YAML.

Kaikki merkintäkielet ovat rakenteista tekstiä, eli ne voidaan hyväksyä pitkäaikais-säilytettäväksi vähintään tekstitiedostoina. Niiden käsittely on kuitenkin monin tavoin pelkkää tekstiä kätevämpää, joten käyttäjiä on hyvä kannustaa merkintäkielien käyttöön tarjoamalla niille PAS-ratkaisussa tekstitiedostoja edistyneempi tuki. Standardinmukaisuus voidaan koneellisesti validoida, ja validoinnin läpäisseiltä tiedostoilta ei tarvitse vaatia yksityiskohtaista rakenteen kuvaavaa dokumentaatiota kuten muiden rakenteisten tekstitiedostojen osalta. Metatietojen tallentamiseen voidaan luoda esimerkiksi tieteenalakohtaisia XML-skeemoja tai JSON-pohjia, sekä tarjota kuvailupalvelussa käyttöliittymä niiden syöttämiseen.

Kansainvälisesti HTML ja XML on laajasti hyväksytty säilytyskelpoisiksi tiedostomuodoiksi. SGML:ää tukevat DANS ja LoC. JSON on mainittu LoC:n suosittelemissa muodoissa, osajoukko JSON-LD myös DANSin suosituksissa. Kaikki merkintäkielet voidaan joka tapauksessa säilyttää ainakin tekstitiedostoina, joita tukevat kaikki organisaatiot.

Tilasto- ja taulukkolaskentaohjelmien tiedostomuodot

Tilasto-ohjelmia käytetään erityisesti yhteiskuntatieteellisessä tutkimuksessa. Kullakin ohjelmalla on tyypillisesti oma tiedostomuotonsa, joista suurin osa on suljettuja. Yksi yleisimmistä tilasto-ohjelmista on kaupallinen SPSS, jonka käyttämät tiedostomuodot SAV ja SPSS Portable ovat muodostuneet alalla de facto -standardeiksi. Suurin osa ohjelmista tukee niitä ainakin osittain, mukaan lukien avoimen lähdekoodin PSPP. Kumpikaan muodoista ei kuitenkaan ole avoimesti dokumentoitu, jälkimmäisen sana "portable" tarkoittaa vain siirrettävyyttä eri tietokonearkkitehtuurien välillä. Toinen paljon erityisesti terveystieteiden alalla käytetty ohjelma on SAS, joka käyttää omaa suljettua tiedostomuotoaan.

Tilasto-ohjelmilla analysoitua dataa on mahdollista muuntaa KDK:ssa säilytyskelpoiseksi hyväksyttyihin taulukkolaskentaohjelmien tiedostomuotoihin tai CSV-muotoon. Muunnoksessa kuitenkin usein häviää informaatiota, eikä tiedostojen avaaminen uudelleen tilasto-ohjelmalla tehtävää jatkokäsittelyä varten aina onnistu ongelmitta. Suomessa Yhteiskuntatieteellinen tietoarkisto FSD on valinnut säilyttämiensä aineistojen muodoksi SPSS Portablen. Muuntaminen muiden tilasto-ohjelmien käyttämistä muodoista siihen tehdään varta vasten muunnoksia varten kehitetyillä kaupallisilla ohjelmilla. SPSS Portable-muoto on käytännön testeissä osoittautunut hyvin alas- ja ylöspäin yhteensopivaksi. Muotoa voi FSD:n mukaan siten suositella säilytysmuodoksi, analyysikäytössä siinä on muutamia rajoituksia tilasto-ohjelmien natiivimuotoihin verrattuna.

Koska FSD aktiivisesti seuraa SPSS Portable -muodon käytettävyyttä ja on tarvittaessa valmis muuntamaan tiedostot tulevaisuudessa paremmin tuettuihin muotoihin, on syytä harkita sen osalta poikkeusta yleisiin kriteereihin, jotka edellyttävät säilytykseen vastaanotettavien tiedostomuotojen määritysten saatavuutta. Muotoa ollaankin KDK:ssa alustavasti hyväksymässä säilytyskelpoiseksi tietyin reunaehdoin. Datan tallentaminen SPSS:n Portable -muodon rinnalle CSV-muotoon on myös mahdollinen vaihtoehto. Tiedostot ovat tyypillisesti kooltaan pieniä, joten siltä osin rinnakkainen tallennus kahteen eri muotoon ei ole ongelma. Ohjelmointitaitoisten tutkijoiden keskuudessa on yleistynyt tilastollinen analyysi avoimen lähdekoodin R-ohjelmistolla. Analysointi tehdään R:n omalla ohjelmointikielellä eikä graafista käyttöliittymää hyödyntäen kuten SPSS:ssä ja monissa muissa tilasto-ohjelmissa. R tukee useita, sekä avoimia että suljettuja tiedostomuotoja. Mm. CSV on yleisesti käytetty, ja ohjelmointikomennot tallennetaan rakenteiseen tekstitiedostoon.

Taulukkolaskentaohjelmia käytetään yleisesti monilla eri tieteenaloilla. Kaksi ylivoimaisesti suosituinta ohjelmaa ovat Microsoft Excel ja LibreOffice/OpenOffice Calc, joilla kummallakin on omat tiedostomuotonsa. LibreOfficen Open Document Spreadsheet (ODS) -muoto on hyväksytty KDK:ssa säilytyskelpoiseksi ja Excelin Office Open XML (XLSX) siirtokelpoiseksi tiedostomuodoksi. Kummankin muodon tuen osalta on kuitenkin huomioitava, että tutkimuksessa käytettäneen kulttuuriaineistoja useammin edistyneille käyttäjille tarkoitettuja ominaisuuksia. Tästä johtuen tiedostojen avaaminen jossain muussa kuin alun perin niiden tuottamiseen käytetyssä ohjelmassa, tai muuntaminen toiseen muotoon ei välttämättä aina onnistu ongelmitta.

Kansainvälisesti CSV, ODS ja XLSX on hyväksytty joko säilytys- tai siirtokelpoisiksi kaikissa tarkastelluissa organisaatioissa. SPSS Portable on hyväksytty säilytettäväksi muodoksi DANSissa ja UKDA:ssa. Myös muutamia muita tilasto-ohjelmien suljettuja tiedostomuotoja tuetaan samoissa kahdessa organisaatiossa vähintään siirtokelpoisina.

Tietokoneavusteisen mallinnuksen tiedostomuodot

Kaksi- ja kolmiulotteista tietokoneavusteista mallinnusta voidaan käyttää eri tutkimusaloilla, ja erityisesti kolmiulotteinen (3D) mallinnus on yleistymässä. Mallit voivat liittyä mittalaitteisiin tai tutkittaviin materiaaleihin, mutta myös esimerkiksi yhteiskuntatieteelliseen tutkimukseen, jossa tutkitaan esineiden tai ympäristön vaikutusta koehenkilöihin. Laadittujen mallien säilyttäminen saattaa olla hyödyllistä joko ymmärrettävyyden säilyttämistä tai uudelleenkäyttöä varten.

Suosittuja mallinnusohjelmia ovat mm. kaupalliset AutoCAD, SolidWorks ja SketchUP sekä avoimen lähdekoodin Blender. Yksinkertaisia 2D-malleja laaditaan usein myös vektorigrafiikan piirtoon soveltuvilla yleisohjelmistoilla kuten Microsoft PowerPoint, LibreOffice Draw, Corel Draw tai Adobe Illustrator. Kolmiulotteisia rakenteita saatetaan varsinaisten mallinnusohjelmien lisäksi muodostaa mittalaitteiden, kuten magneettikuvauslaitteiden (MRI) tai 3D-skannerien tuottamien kuvien pohjalta.

Mikäli tavoitteena on vain ymmärrettävyyden säilyttäminen, esimerkiksi käytetyn mittalaitteen kuvaaminen, laaditut mallit voi tulostaa kuvina PDF-tiedostoiksi, joiden säilytykseen KDK:ssa on jo varauduttu. Mallien muokkaukseen tai muuhun uudelleenkäyttöön PDF-muoto ei kuitenkaan sovellu.

Yleiskäyttöisten vektorigrafiikkaohjelmistojen tiedostomuodoista on KDK:ssa siirto- tai säilytyskelpoisiksi hyväksytty Microsoft PowerPoint ja LibreOffice Draw -ohjelmien tiedostomuodot. Ne eivät kuitenkaan juurikaan sovellu uudelleenkäyttöön mallintamisen, varsinkaan 3D-mallien osalta.

Mallinnusohjelmien tiedostomuodoista eniten käytetty on AutoCADin DWG-muoto. Sen kehitys on Autodesk-yrityksen kontrolloimaa, eikä virallista dokumentaatiota ole julkisesti saatavilla. Open Design Alliance on kuitenkin kuvannut DWG-muodon varsin tarkasti [ODA_DWG_Specification] ja se on melko hyvin tuettu useissa eri ohjelmissa. Mikäli DWG-muodon hyväksymistä joko siirto- tai säilytyskelpoiseksi tiedostomuodoksi harkitaan, validointi sekä hyväksymisvaatimukset on määriteltävä tämän epävirallisen dokumentaation pohjalta.

Muita varteenotettavia 2D- ja 3D-mallinnuksen tiedostomuotoja ovat mm. 3D Studio (3DS), AutoCAD Drawing Interchange Format (DXF), Blenderin käyttämä BLEND, Initial Graphics Exchange Specification (IGES), Product Representation Compact (PRC), STEP File, Wavefront OBJ ja X3D. Näistä STEP (ISO 10303-21) ja IGES (v. 5.3, ANSI 1996) ovat molemmat virallisia standardeja ja hyvin dokumentoituja, mutta ominaisuuksiltaan vanhentuneita. X3D on uudempi, erityisesti 3D-sisällön esittämiseen verkossa kehitetty standardi, joka ei kuitenkaan erityisen hyvin sovellu mallien tallennukseen uudelleenkäyttöä varten.

Autocad DXF on DWG-muodon kehittäneen Autodeskin tarjokas tiedonsiirtomuodoksi eri CAD-ohjelmien välillä. Se on, toisin kuin DWG, avoimesti dokumentoitu, mutta ei tue kaikkia uusimpia ominaisuuksia. 3D Studio on saman yrityksen erityisesti 3D-mallinnukseen kehittämä muoto, joka dokumentaation puutteista huolimatta on DWG:n tapaan noussut de facto -standardiksi. Blenderin käyttämä BLEND-tiedostomuoto on avoimen lähdekoodin taustansa myötä avoimesti dokumentoitu sekä monipuolinen, mutta rakenteeltaan poikkeuksellinen ja muissa ohjelmissa huonosti tuettu. Wavefront OBJ on dokumentoitu, varsin yksinkertainen muoto 3D-rakenteiden esittämiseen. ISO-standardoitu PRC (ISO 14739-1:2014) on tarkoitettu 3D-mallien upottamiseen PDF-tiedostoihin. Se ei kuitenkaan ole osa KDK:ssa hyväksyttyjä PDF 1.7 tai PDF/A-standardeja.

Mikään edellä kuvatuista 2D- ja 3D-mallinnuksen tiedostomuodoista ei ole pitkäaikaissäilytyksen kannalta erityisen hyvä. Joko dokumentaatiossa tai yhteensopivuudessa on puutteita, muodot ovat vanhentuneita tai ne soveltuvat vain ymmärrettävyyden säilytykseen eli mallien esittämiseen, eivät uudelleenkäyttöön. On myös toistaiseksi epäselvää, minkä verran 2D- ja 3D-malleja tutkimusaineistoissa käytetään, ja mitkä muodot ovat yleisimpiä.

Kansainvälisesti neljä tarkastelluista kuudesta organisaatiosta (DANS, LAC, NAA, UKDA) hyväksyy AutoCADin DWG- ja DXF-muodot siirto- tai säilytyskelpoisina. Suositelluin muoto on DANS:n mukaan DXF, UKDA:n listalla puolestaan DWG.

Geenisekvenssien tiedostomuodot

Geenisekvenssien tallennukseen käytetään yleisesti BAM/SAM- ja CRAM-tiedostomuotoja, jotka on esitelty liitteessä C. Näiden lisäksi varteenotettavia muotoja ovat BCF/VCF sekä FastQ.

Sekvensoijan tuottama raakadata tallennetaan tyypillisesti FastQ-muodossa, ja siitä edelleen prosessoitu data BAM-muodossa. BAM-muotoa voi kuitenkin käyttää myös FastQ:n korvaajana, ja sen rakenne mahdollistaa FastQ:ta monipuolisemman metatietojen tallennuksen. FastQ on BAM:n tapaan avoin, dokumentoitu tiedostomuoto, jonka etu on yksinkertaisuus. Paremmin suunnitellun metatietojen tallennuksen ansiosta BAM lienee kuitenkin pitkäaikaissäilytyksen kannalta parempi valinta.

CRAM-muoto on otettu käyttöön tilan säästämiseksi — se on käytännössä BAM-tiedosto, josta on kontrolloidusti ja dokumentoidusti jätetty pois osa geenisekvenssin informaatiosta. Lisäominaisuudet tekevät CRAM:sta BAM-muotoa monimutkaisemman. Koska geenitutkimuksen aineistot ovat suuria, jopa kymmeniä tai satoja teratavuja, on CRAM-muodon tukeminen tästä huolimatta PAS-ratkaisussa perusteltua.

Variant Call Format (VCF) -muotoa sekä sen binäärimuotoa BCF:ää käytetään, kun genomista on prosessoitu tietoa. VCF/BCF-tiedosto voi sisältää yhden tai useamman ihmisen genomin, ei enää puhdasta sekvenssiä vaan genotyyppejä. Kyseessä on suhteellisen uusi, mutta käytännössä jo de facto -standardiksi muodostunut tiedostomuoto. Se täydentää käyttötarkoituksensa osalta BAM- ja CRAM-tiedostomuotoja, on avoimesti dokumentoitu ja soveltuu siten myös PAS-ratkaisussa tuettavaksi muodoksi.

Tarkasteltujen ulkomaisten säilytysorganisaatioiden tiedostomuotojen listoilla geenisekvenssien tiedostomuotoja ei ole lueteltu. Aineistoja säilytetään tyypillisesti erityisesti geenitutkimukseen keskittyvissä tietopankeissa, joita kansainvälinen tutkimusyhteisö käyttää aktiivisesti. Sen myötä alan tiedostomuodot ovat melko hyvin vakiintuneet.

Aivotutkimuksen tiedostomuodot

Aivojen toimintaa tutkitaan tyypillisesti käyttäen magneettikuvauksella (MRI) saatuja kuvasarjoja. Muita usein käytettyjä tekniikoita ovat aivosähkökäyrät (elektroenkefalografia, EEG) ja aivomagneettikäyrät (magnetoenkefalografia, MEG).

MRI-tekniikan avulla saadaan tietoa sekä aivojen anatomiasta että toiminnasta (functional MRI), joita voidaan verrata eri koehenkilöiden kesken ja eri koetilanteissa. Dataa kertyy helposti varsin paljon. Aivotutkimus ja MRI on hyvä esimerkki siitä, miten tiedostomuodot ja yhteiset käytännöt ovat kehittyneet eri tutkimusryhmien välisen yhteistyön lisääntyessä sekä teknisten edistysaskelien mukanaan tuomien uusien tarpeiden myötä.

MRI-kuvauslaitteet tuottavat tyypillisesti DICOM-tiedostomuodossa olevia kuvatiedostoja, jotka sisältävät myös kuvauksessa käytetyt parametrit ja muita metatietoja. Parametrit ovat kuitenkin valmistajakohtaisia, ja lisäksi DICOM-standardi tarjoaa mahdollisuuden liittää tiedostoon suljettuja osia, joita ei esimerkiksi tekijänoikeussyistä saa levittää edelleen. Siksi DICOM-tiedostot usein muunnetaan laiteriippumattomaan NIfTI-muotoon, joka on esitelty liitteessä C. Samassa liitteessä esitelty, alan tutkimusaineistoissa yleiseksi käytännöksi muodostunut BIDS-hakemistorakenne edellyttää NIfTI-muodon käyttöä. BIDS määrittelee myös tiedostojen nimeämiskäytännön sekä metatietojen tallennuksen TSV- ja JSON-tiedostoihin.

NIfTI-muoto ei kuitenkaan yksinään täytä kaikkien aivotutkijoiden tarpeita. Uusien menetelmien, joissa mm. vertaillaan aivoissa kulkevia signaaleita, pinnanmuotoja sekä aivojen eri osien keskinäistä vaikutusta toisiinsa, tuottamaa dataa ei kaikilta osin voida tallentaa NIfTI-muotoon. Tämä on johtanut GIFTI- ja CIFTI-tiedostomuotojen määrittelyyn. GIFTI-tiedostoihin tallennetaan pintoihin liittyvää dataa, CIFTI puolestaan on lisämääritys, jonka mukaisesti NIfTI-tiedostoihin tallennetaan lisää metatietoja XML-muodossa sekä täydentävää mittausdataa. Kumpikaan niistä ei ole vielä yhtä laajalle levinnyt kuin NIfTI, eivätkä ne toistaiseksi sisälly BIDS-määritykseen. Tiedostomuotojen käyttöönotto alalla tunnetussa ja arvostetussa Human Connectome -projektissa [Human_Connectome] tuo ne kuitenkin osaksi yhä useampien aivotutkijoiden aineistoja.

Tiedostomuodot vaikuttavat asiallisesti dokumentoiduilta, mutta toisaalta olevan vielä kehitysvaiheessa. Tähän viittaa uudehkojen CIFTI- ja GIFTI-muotojen lisäksi se, että sekä NIfTI- että CIFTI-muodoista on viimeisen kolmen vuoden sisällä julkaistu uusi versio (NIfTI-2 ja CIFTI-2), joista kumpikaan ei ole edellisen version kanssa täysin yhteensopiva. Alalla vallitseva sisäinen pyrkimys aineistojen uudelleenkäytettävyyteen johtanee muotojen vakiintumiseen vähitellen. Tämä lienee suuntaus myös EEG- ja MEG-tiedostojen osalta, vaikka niissä ei olla tähän mennessä saavutettu yhtä suurta yhtenäisyyttä kuin MRI-kuvien tallennusmuodoissa.

Pitkäaikaissäilytyksen näkökulmasta mainitut aivotutkimuksen tiedostomuodot ovat avoimuuden, dokumentaation ja ohjelmistotuen osalta hyväksyttäviä. Metatietojen osalta on kuitenkin määriteltävä, mitkä vapaaehtoiset kentät edellytetään täytettäväksi säilytettäviksi hyväksyttävissä tiedostoissa, sekä niiden sisältöön liittyvät yksityiskohdat. Lisäksi on varauduttava siihen, että tiedostoja joudutaan myöhemmin muuntamaan uudempiin tiedostomuotoihin, jotta ne pysyvät alan nopeahkossa kehityksessä mukana.

Projektissa tarkastelluista kuudesta ulkomaisesta säilytysorganisaatiosta yksikään ei mainitse suositeltujen tiedostomuotojen listalla aivotutkimuksen tiedostomuotoja. Geenisekvenssiaineistojen tapaan niitä tallennetaan pääasiassa tutkimusalan omiin palveluihin, jotka samalla kontrolloivat myös tiedostomuotojen kehitystä.

Lääketieteellisen tekniikan tiedostomuodot

Erikseen esitellyn aivotutkimuksen lisäksi on monia muita lääketieteellistä tekniikkaa hyödyntäviä tutkimusaloja. Niissä käytetään tyypillisesti kalliita mittalaitteita, joiden yksityiskohdat ovat usein laitevalmistajien liikesalaisuuksia. Monet laitteet noudattavat DICOM-standardia, joka määrittelee sekä laitteiden välisen liitäntäprotokollan että kuvatiedostomuodon. DICOM itsessään on siis hyvin dokumentoitu, mutta tietyt osat DICOM-tiedostoista sekä muut laitteiden tuottamat tiedostomuodot ovat usein valmistajakohtaisia, niiden dokumentaatio ei ole avoimesti saatavilla ja tiedostojen käsittelyyn tarvitaan suljetun lähdekoodin erikoisohjelmisto.

Pitkäaikaissäilytyksen kannalta lääketieteellisen tekniikan ala on haastava. Läheskään kaikki alan tutkimuksen osa-alueet eivät ole toistaiseksi ryhtyneet tiedostomuotojen yhtenäistämiseen aivotutkijoiden tapaan. Aineistojen uudelleenkäyttöä rajoittavat usein myös yksityisyyden suojaan liittyvät seikat. DICOM-kuvatiedostojen tukea kannattaa kuitenkin harkita.

Kansainvälisesti DICOM on hyväksytty DANSin ja LAC:n listoilla, muiden tarkasteltujen organisaatioiden listoilta sitä ei löydy.

Kielitieteen aineistojen tiedostomuodot

Kielitieteessä käytetään erityyppisiä aineistoja, joista kuhunkin liittyvät omat tiedostomuotonsa. Kolme pääryhmää ovat tekstimuotoiset aineistot, ääni- ja videotallenteet, joista kaikkiin voi lisäksi liittyä tutkimuksessa tehdyssä analyysissa tuotettuja tietoja.

Tekstimuotoisten aineistojen analyysissa käytetään enimmäkseen rakenteisia tekstitiedostoja. Rakenne voi sisältää esimerkiksi lauseoppiin, muoto-oppiin eli morfologiaan sekä semantiikkaan liittyviä merkintöjä, vaikkapa analysoidun tekstin sanojen erittelyn lauseenjäseniksi ja sanojen taivutusmuodot kuvaavat merkinnät, kukin omissa sarakkeissaan. Tiedostomuodot ovat yleensä periaatteessa avoimia, mutta eivät läheskään aina hyvin dokumentoituja. Eräänlaiseksi de facto -standardiksi on muodostunut CoNLL-U-muoto [CoNLL-U], joka erottuu dokumentaationsa osalta edukseen muista samankaltaisista muodoista.

Ymmärrettävyyden säilyttämisen kannalta olennaisten metatietojen esittämiseen ei suurimmassa osassa tiedostomuodoista ole mahdollisuutta tai yksikäsitteistä paikkaa. Myöskään CoNLL-U-tiedostoissa ei ole minkäänlaista header-osiota tai muuta tapaa merkitä metatietoja. Teknisellä tasolla mm. eri merkistöjen käyttö saattaa johtaa yhteensopivuusongelmiin erityisesti vanhempien aineistojen kanssa. Uusissa aineistoissa käytetään lähes poikkeuksetta UTF-8-merkistöä. Kuvailun tasolla esimerkiksi tekstin lähde, konteksti ja käytetty kieli ovat ymmärrettävyyden kannalta olennaisia metatietoja. Metatiedot voidaan tallentaa erilliseen tiedostoon esimerkiksi XML- tai JSON-muodossa. CLARIN-hankkeessa erilaisia metatietoskeemoja ja -formaatteja hallinnoidaan Component MetaData Infrastructure -rakenteen avulla [CLARIN_CMDI].

Suomi24-esimerkkiaineistossa käytetty VRT-muoto on eräänlainen sekamuoto, jossa XML:n kaltaisesti tallennettuja metatietoja on yhdistetty samaan tiedostoon CoNLL-U-tyyppisten analyysimerkintöjen kanssa. Kyseessä ei kuitenkaan ole XML-tiedosto, ja rakenne sekä lyhenteet on CoNLL-U:ta huonommin dokumentoitu.

Text Encoding Initiative (TEI) on sekä yhteisö että kyseisen yhteisön kehittämä XML-pohjainen standardi tekstimuotoisten aineistojen tallentamiseen. Se mahdollistaa sekä alkuperäisen tekstin, rakenteeseen liittyvien merkintöjen että metatietojen tallentamisen yhteen tiedostoon. TEI-standardi on hyvin laaja, mutta laadittu joustavasti siten että määrityksistä voidaan hyödyntää vain kulloinkin tarvittavat osat. Teksti voidaan TEI:n mukaisesti tallentaa

lähes sellaisenaan, esimerkiksi eritellen XML-tageilla vain kappaleet samaan tapaan kuin HTML-dokumenteissa, tai rikastaa sitä hyvinkin yksityiskohtaisilla merkinnöillä kuhunkin yksittäiseen sanaan liittyen.

TEI-muoto on XML-pohjaisuutensa ansiosta koneluettava ja soveltuu pitkäaikaissäilytettäväksi. Myös syntaksin sekä TEI-skeeman mukaisuuden tarkistavia validaattoreita on valmiiksi saatavilla. Pitkäaikaissäilytyksen osalta on syytä määritellä, mitkä metatietokentät vaaditaan täytettäviksi, sekä laajentaa validointia kattamaan niiden sisällön tarkistus.

TEI-standardin joustavuuden ansiosta lähes kaikki kielitieteen rakenteiset tekstitiedostot voitaisiin periaatteessa muuntaa TEI-muotoon. Monet valmiit analyysityökalut eivät kuitenkaan tue sitä ja itse ohjelmoivat kielitieteilijät suosivat usein yksinkertaisempia muotoja kuten CoNLL-U. Siten niiden tukeminen on myös perusteltua, käyttäen samoja kriteerejä dokumentaation ja metatietojen suhteen kuin yleisesti rakenteisten tekstitiedostojen kohdalla.

Ääni- ja videotallenteet käyttävät samoja tiedostomuotoja kuin KDK:ssa, ja niiden osalta voidaan soveltaa KDK:ssa jo laadittuja määrityksiä. Kielitieteessä on kuitenkin tärkeää pystyä tekemään merkintöjä, jotka viittaavat tiettyihin ajankohtiin tallenteessa. Merkinnät tallennetaan tyypillisesti omaan erilliseen tiedostoonsa, ja niitä varten on kehitetty ELAN Annotation Format (EAF) -muoto. Kyseessä on melko yksinkertainen XML-pohjainen muoto, joka soveltuu hyvin pitkäaikaissäilytettäväksi. Kuten yleisesti uusien tiedostomuotojen kohdalla, pitkäaikaissäilytyksen osalta on syytä määritellä vaadittavat metatietokentät ja niiden yksityiskohdat. Lisäksi on varmistettava, että EAF-tiedosto ja siihen liittyvä ääni- tai videotallenne pysyvät yhtenä kokonaisuutena.

Kansainvälisesti TEI-muoto löytyy CINESin, DANSin, ja LoC:n hyväksyttyjen muotojen listoilta. Muita kielitieteen tiedostomuotoja ei tarkastelussa mukana olleilta listoilta löydy. XML-pohjaiset muodot on kuitenkin hyväksytty yleisesti kaikissa, ja niiden hyväksyntään on osassa organisaatioista tarjolla myös tarkempia ohjeita.

Seismologian tiedostomuodot

Seismologian yleisin tiedostomuoto mittausdatan tallennuksessa on FIRE-esimerkkiaineistossakin käytetty SEG-Y-muoto. Toinen laajasti käytetty on Seismic Unix, joka on saman nimisen, suositun avoimen lähdekoodin analyysiohjelmiston tiedostomuoto. Seismic Unix -tiedostot ovat helposti muunnettavissa SEG-Y-muotoon ja päinvastoin, joten pitkäaikaissäilytyksen kannalta lienee riittävää tukea SEG-Y-muotoa. Sitä pystyvät sekä lukemaan että kirjoittamaan kaikki tärkeimmät alan ohjelmistot.

Datatiedostojen lisäksi seismologian aineistojen tulkinnassa olennaisia tietoja ovat havaintopisteiden koordinaatit, mittauksissa käytetyt parametrit, havaintopäiväkirja sekä kenttätyöraportti, joka sisältää käytettyjen parametrien lisäksi mittauksen sanallisen kuvailun. Näiden tietojen tallentamiseen ei ole yleisesti hyväksyttyä käytäntöä. Joitakin parametreja voidaan tallentaa SEG-Y-tiedostojen header-osioon, mutta havaintopisteiden koordinaatit, havaintopäiväkirja sekä kenttätyöraportti ovat tyypillisesti rakenteisia tekstitiedostoja tai tekstinkäsittelyohjelmalla tuotettuja dokumentteja. Niiden pitkäaikaissäilytyksen tuessa tulee käyttää yleisiä rakenteisten tekstitiedostojen kriteerejä mm. dokumentaation osalta. Erityisesti on syytä huomioida maantieteellisten koordinaattien yhteensopivuus muiden aineistojen kanssa, ja tarvittaessa muuntaa havaintopisteiden koordinaatit johonkin PAS-ratkaisussa tuettavaksi valittuun koordinaattijärjestelmään.

Seismologian tiedostomuotoja ei löydy tarkasteltujen ulkomaisten säilytysorganisaatioiden hyväksyttyjen muotojen listoilta.

Ilmakehätieteen ja ekosysteemien tutkimuksen tiedostomuodot

Ilmakehätieteen ja ekosysteemien, tai laajemmin maapallon luonnonjärjestelmän (Earth System) tutkimuksessa käytetään tyypillisesti maantieteellisesti hajautetusti sijoitettuja mittalaitteita. Hankkeet ovat usein kansainvälisiä, mikä vaikuttaa myös datan keruuseen ja käsittelyyn.

Tavallisimmat tiedostomuodot ovat rakenteinen teksti, CSV ja HDF5. Referenssinä käytetään lisäksi kaukokartoitusdataa, joka on useimmiten GeoTIFF- tai HDF5:een pohjautuvassa NetCDF-muodossa. Tiedostomuodot ovat pääsääntöisesti avoimia ja hyvin dokumentoituja.

Myös tietokantojen käyttö on alalla yleistä, erityisesti kansainvälisissä hankkeissa. Useimmiten tietokannat eivät suoraan korvaa mittalaitteiden tuottamia datatiedostoja, vaan täydentävät niitä ja tarjoavat tutkijoille aineistojen käyttöä helpottavia rajapintoja. Tietokantoja ylläpitävät kansainväliset infrastruktuurit keskittyvät usein tiettyihin yksittäisiin suureisiin, ja keräävät niiden mittaustuloksia useilta tutkimusryhmiltä eri puolilta maailmaa. Suomalainen SMEAR-hanke lähettää dataa useampaan eri kansainväliseen infrastruktuuriin, ja ylläpitää lisäksi enemmän suureita sisältävää, mutta maantieteellisesti rajatumpaa aineistokokonaisuutta Suomessa [SMEAR_AVAA].

Tutkimuksessa käytetään tyypillisesti aineistoja useammista eri lähteistä. Siihen liittyen on huomattava, että tallennuskäytännöt eri aloilla poikkeavat jonkin verran toisistaan. Ekosysteemeissä data on yleensä suurepohjaista, jolloin ilmoitetaan esimerkiksi lämpötila riippumatta siitä, millä laitteella se on mitattu, ja mittalaitetta voidaan vaihtaa kesken datan keruun. Ilmakehätieteissä puolestaan aloitetaan uusi datasetti mittalaitteen vaihtuessa.

Ilmakehätieteen ja ekosysteemien tutkimuksen tiedostomuotoja ei ole tarkasteltujen ulkomaisten organisaatioiden listoilla mainittu erikseen. CSV on kuitenkin hyväksytty joko säilytys- tai siirtokelpoiseksi kaikissa (CINESissä tosin vain tekstinä ilman varsinaista CSV-tukea), GeoTIFF kaikissa paitsi NAA:ssa ja HDF5 kolmessa (CINES, DANS, LoC) tarkastelluista organisaatioista.

Avaruustutkimuksen tiedostomuodot

Avaruustutkimuksessa käytetään monenlaista havaintodataa, mm. teleskoopeilla otettuja kuvia kohteista (Maa, Aurinko, muut planeetat ja tähdet) eri aallonpituuksilla ja satelliittien ympäristödataa, joka voi olla mittauksia esim. satelliitin plasmaympäristöstä (tiheys, lämpötila, virtausnopeus), sen ympärillä vallitsevasta sähkömagneettisesta kentästä tai satelliitin kohtaamasta säteilystä. Kullakin havaintotyypillä on käytössä useita erilaisia tiedostomuotoja.

Hiukkassäteilydataa jaetaan yleensä joko CDF- tai tekstimuotoisena, joskus myös HDF-5 -muotoisena. Tekstitiedostotyypeistä yleisin on CSV, myös rakenteisia tekstitiedostoja vakiolevyisin sarakkein käytetään. Plasma- ja kenttädataa jaetaan CDF- tai tekstimuodossa.

Tähtitieteessä ja satelliittikuvien jakelussa yleisin tiedostomuoto on FITS, joka on myös Planck-esimerkkiaineiston käyttämä muoto. Se on rakenteeltaan melko monimutkainen muoto, joka mahdollistaa kuvien lisäksi hyvin monenlaisen datan tallentamisen. FITS on avoin, dokumentoitu ja soveltuu pitkäaikaissäilytettäväksi kunhan vaadittavat tekniset metatiedot on määritelty. Tarkempi kuvaus muodosta on liitteessä C. Ihmissilmin tarkasteltavaksi tarkoitettuja satelliittikuvia levitetään myös yleisissä kuvatiedostomuodoissa kuten TIFF, PNG tai JPEG, jotka on KDK:ssa hyväksytty säilytyskelpoisiksi.

Kansainvälisesti HDF5 on hyväksytty siirto- tai säilytyskelpoiseksi kolmessa (CINES, DANS, LoC) kuudesta tarkastellusta organisaatiosta. LoC mainitsee lisäksi CDF:n. FITS-muotoa ei löydy yhdenkään näiden organisaatioiden listoilta, mutta se on vakiintunut muoto tieteenalan omissa säilytyspalveluissa. Tekstitiedostot ja yleiset kuvatiedostomuodot ovat laajasti sekä kansallisesti että kansainvälisesti säilytyskelpoisiksi hyväksyttyjä.

Hiukkasfysiikan ja ydinfysiikan tiedostomuodot

Hiukkasfysiikan ja ydinfysiikan tutkimuksessa käytetään tyypillisesti kalliita, nimenomaan alan tutkimukseen suunniteltuja mittalaitteita sekä pitkälle erikoistuneita ohjelmistoja. Tiedostomuodot ovat usein ohjelmistokohtaisia, mutta melko vakaita, koska tutkimusprojektit ovat pitkäkestoisia ja dataa analysoidaan jopa useiden kymmenien vuosien ajan. Ohjelmistojen lähdekoodi on useimmiten saatavilla ja tiedostomuodot yleensä periaatteessa avoimia, mutta niiden dokumentaatiossa on puutteita. Aineistot eivät siksi ole kovin helposti siirrettävissä ohjelmasta toiseen.

Tunnetuin alalla käytetty tiedostomuoto on CERNissä kehitetty, samannimisen ohjelmiston mukaan nimetty ROOT-tiedostomuoto [ROOT]. Se on optimoitu erityisesti suurteholaskennan tarpeisiin, koska datamäärät ovat erittäin suuria ja analyysi vaatii runsaasti laskentakapasiteettia. Itse tiedostomuoto on varsin hyvin dokumentoitu, mutta sen käsittelyyn tarkoitettu ROOT-analyysiohjelmisto ja -ohjelmointikirjasto on laaja ja monimutkainen.

Muita alalla yleisesti käytettyjä tiedostomuotoja ovat mm. RadWare, MED ja ENDSF. Näistä kaksi ensimmäistä ovat ensisijaisesti analyysiohjelmistoja ja itse tiedostomuotojen dokumentaatio on puutteellinen. RadWaren usein esitettyjen kysymysten listan vastaus ohjelman käyttämien tiedostomuotojen rakenteesta kuvaa tilannetta: "Muotoja on useita ja ne ovat erilaisia. Paras ja täsmällisin tapa saada tietoa niistä on katsoa ohjelman lähdekoodia tiedostoja lukevien ja kirjoittavien osioiden kohdalta" [RADWARE_FAQ]. ENDSF on selkeämmin nimenomaan tiedostomuoto, joka on myös asiallisesti dokumentoitu. Myös tietokantoja on käytössä jonkin verran, mm. Yhdysvaltojen National Nuclear Data Centerin ylläpitämänä [NNDC_Databases].

Suurin osa hiukkas- ja ydinfysiikan tiedostomuodoista on alan tutkimusorganisaatioissa itse kehitettyjä. Myös raakadatan tallennus sekä pitkäaikainen säilytys on usein keskitetty samoihin organisaatioihin. Pitkäaikaissäilytyksen osalta on siten olennaista ensinnäkin selvittää, minkä aineistojen tallennus Suomessa ylipäätään tuo tutkijoille lisäarvoa, ja sen perusteella katsoa, mitä tiedostomuotoja PAS-ratkaisun tulisi tukea sekä mitä metatietoja niiden yhteyteen tarvitaan. Valittujen tiedostomuotojen osalta on huolehdittava asianmukaisesta dokumentaatiosta.

Hiukkas- ja ydinfysiikan tiedostomuotoja ei löydy projektissa tarkastellun kuuden ulkomaisen säilytysorganisaation listoilta. Aineistot säilytetään tyypillisesti hiukkas- ja ydinfysiikan tutkimukseen keskittyvissä organisaatioissa.

6.4 Tietokantamuotoiset aineistot

Tutkimusaineistoja tallennetaan yhä useammin tietokantoihin, joista haluttuja osia aineistoista voidaan hakea ja ladata perinteisiä tiedostoja joustavammin. Erityisesti laajat, kansainvälisesti hyödynnetyt aineistokokonaisuudet hyödyntävät tietokantoja. Joko koko aineisto voidaan tallentaa tietokannan sisään tai hyödyntää kantaa indeksinä, jolloin itse data on perinteisesti tiedostoissa ja kannan avulla voidaan hakea ja valita halutut tiedostot. Myös näiden kahden toimintamallin yhdistelmä on mahdollinen. Tietokantaan tallennettua aineistoa voidaan rajapintojen kautta ladata eri tiedostomuodoissa, joita voidaan joustavasti muuttaa tarpeen mukaan.

Tietokannan rakenteen monimutkaisuus vaikuttaa olennaisesti siihen, kuinka vaativaa sen säilytys on. Koko ei välttämättä kerro paljoa: suurikin tietokanta voi olla rakenteeltaan yksinkertainen tai pieni kanta sisältää paljon erilaisia tauluja tai objekteja sekä niiden välisiä suhteita. On myös huomattava, että kantoihin voidaan tallentaa hyvin monenlaista sisältöä, mukaan lukien binäärisiä objekteja. Tietokannan säilytyskelpoisuutta arvioitaessa on huomioitava itse kannan lisäksi kaikki sen sisältämät tietotyypit.

Automaattisten validointityökalujen tarve korostuu tietokantojen säilytyksessä. Tietokantoja ei voi vaikkapa teksti- ja kuvatiedostojen tapaan avata ohjelmassa ja tarkastaa silmämääräisesti. Niitä ei myöskään voi säilyttää suoraan samassa muodossa kuin missä niitä käytetään, vaan sisältö on luettava tietokantapalvelimelta erilliseen säilytysmuotoon. Monille tietokannoille on saatavilla visualisointityökaluja, joiden avulla voidaan selata kannan tietoja sekä esittää sen rakenne, mutta tietosisällön täydellisyyden sekä säilytysmuodon oikeellisuuden varmistamisen tulee perustua automatiikkaan.

Säilytyspalvelusta ladattujen tietokantamuotoisten aineistojen uudelleenkäyttö on oma haasteensa. Säilytysmuotoon tallennettu aineisto on siirrettävä jälleen tietokantapalvelimelle, jotta monipuolisia ominaisuuksia hakea ja ladata aineistoa voidaan hyödyntää. Palvelinohjelmiston asentaminen on loppukäyttäjälle hankalaa. Tietokannan päälle saattaa lisäksi olla laadittu käyttöliittymä, jonka toimimaan saaminen on vaikeaa. Käyttöliittymät vastaavat säilytyksen ja uudelleenkäytön osalta edellisessä luvussa käsiteltyjä ohjelmien lähdekoodi- ja binääritiedostoja.

Relaatiotietokannat ja SIARD-muoto

Yleisimmin käytetty tietokantatyyppi on relaatiotietokanta, joiden käsittelyyn soveltuvia ohjelmistoja on saatavilla eri valmistajilta. Suosittuja ohjelmistoja ovat mm. IBM DB2, Microsoft SQL Server, MySQL, Oracle ja PostgreSQL. Ne pohjautuvat periaatteessa SQL-standardiin, mutta kullakin valmistajalla on siihen omat laajennuksensa ja poikkeuksensa. Erityisesti kantojen ohjelmointitoiminnot ovat yleensä valmistajakohtaisia eivätkä keskenään yhteensopivia.

Kaikkien tietokantojen sisältö voidaan lukea palvelinohjelmiston mukana tulevien työkalujen avulla varmuuskopioksi ns. dump-tiedostoon, josta tiedot voidaan palauttaa uuteen, tyhjään kantaan. Palauttaminen onnistuu useimmiten saman valmistajan ohjelmiston uudempaan versioon, mutta yhteensopivuudesta erityisesti pitkällä aikajänteellä ei ole takeita. Tämä tekee tietokannoista ja niiden dump-tiedostoista haastavia pitkäaikaissäilytyksen kannalta.

Relaatiotietokantojen säilytystä varten alettiin 2000-luvun alussa Sveitsin kansallisarkistossa kehittää SIARD-muotoa [SIARD_2004]. Sen tavoitteena oli olennaisen tietosisällön säilyttäminen SQL-standardiin tukeutuen, tietokantojen valmistajakohtaisista ratkaisuista ja laajennuksista riippumattomasti. Tiedostomuodossa on lisäksi kentät kuvailevalle ja tekniselle metatiedolle, jotta ymmärrettävyyden säilyminen voidaan taata.

SIARDin versio 1.0 hyväksyttiin Sveitsissä kansalliseksi standardiksi vuonna 2013. Tanskan kansallisarkisto oli kuitenkin jo ottanut käyttöön jonkin verran alkuperäisestä poikkeavan SIARDDK-muodon, ja Portugalin kansallisarkistossa oli kehitetty osin vastaava DBML-muoto. Näistä kolmesta saatujen kokemusten pohjalta kehitettiin SIARD 2.0, joka näyttää olevan vakiintumassa relaatiotietokantojen säilytysmuodoksi. Se hyväksyttiin Sveitsissä kansalliseksi

standardiksi kesäkuussa 2016 [SIARD_Standard], ja on hyväksytty säilytyskelpoiseksi muodoksi myös ranskalaisessa CINESissä ja tanskalaisessa DANSissa.

SIARD 2.0 tukee kaikkia SQL:2008-standardin datatyyppejä, ja arvoille määriteltyjä rajoitteita (constraints). SIARD säilyttää taulujen väliset yhteydet eli relaatiot, jotka häviäisivät mikäli taulut tallennettaisiin erikseen esimerkiksi CSV-muotoisina tiedostoina. Eri relaatiotietokantojen valmistajakohtaiset ominaisuudet, erityisesti ohjelmointitoiminnot, eivät ole tuettuja. Monissa aineistoissa niitä ei kuitenkaan ole käytetty tai ne eivät ole tärkeitä aineiston säilyttämisen kannalta. Asia on syytä kuitenkin varmistaa aineistokohtaisesti ennen siirtoa säilytykseen.

Toteutukseltaan SIARD on XML-muotoinen tiedosto ja se käyttää Unicode-merkistöä, yleensä UTF-8. Se voi kuitenkin sisältää binäärisiä osia, mikäli tietokantaan on tallennettu binäärisiä objekteja (BLOB), esimerkiksi kuvia. SIARDin versio 2.0 tukee binääristen osien tallentamista erillisiin tiedostoihin, jolloin niiden ymmärrettävyyden säilytyksestä voidaan huolehtia erikseen. Itse SIARD-tiedosto ilman binäärisiä osia voitaisiin jo KDK-PAS:n nykymääritysten mukaan säilyttää XML-tiedostona, mutta on parempi määritellä erikseen SIARD-tuki.

SIARD-muotoisia tiedostoja voidaan tuottaa avoimen lähdekoodin Database Preservation Toolkit -ohjelmalla [DBPTK]. Se tukee yleisimpiä käytössä olevia relaatiokantoja, lukien rakenteen ja tietosisällön kannasta ja tallentaen ne SIARD-muotoon. Tulokseksi saadut tiedostot voidaan samalla ohjelmalla muuntaa takaisin joko samaan tai muun valmistajan relaatiotietokantaan.

SIARD-tiedostoista ei voi hakea ja ladata tietoja SQL-kielisillä komennoilla varsinaisten relaatiokantojen tapaan, se on puhtaasti säilytykseen tarkoitettu muoto. Yksikään yleisistä tietokannoista ei myöskään tue tietojen tuomista kantaan suoraan SIARD-muodosta, vaan siihen on käytettävä edellä mainittua muuntotyökalua, joka ei vielä ole kypsä, peruskäyttäjälle soveltuva ohjelma. Nämä seikat hankaloittavat SIARDin käyttöä, vaikka itse tiedostomuoto vaikuttaa hyvin määritellyltä.

SMEAR-esimerkkiaineisto ja sen muuntaminen SIARD-muotoon

Esimerkkiaineistoksi saatu SMEAR-tietokanta on kooltaan suurehko, mutta rakenteeltaan yksinkertainen MySQL-relaatiotietokanta. Tauluja on muutama kymmenen ja niissä on runsaasti sarakkeita, mutta taulut ovat joko itsenäisiä tai niiden väliset suhteet ovat helposti ymmärrettäviä. Binäärisiä osia ei ole eikä MySQL:n ohjelmointitoimintoja ole käytetty. Mikäli tietokannan päälle rakennettua www-käyttöliittymää ei huomioida, aineisto on ainakin periaatteessa helpohko säilytettävä.

Aineiston muuntamista SIARD-muotoon kokeiltiin lyhyesti testiympäristössä. Käyttöjärjestelmä oli Ubuntu Linux 14.04 LTS, tietokanta MySQL:n versio 5.5.50, Java-ympäristön versio 1.7.0_111 (OpenJDK IcedTea 2.6.7) ja Database Preservation Toolkit-ohjelmistosta käytettiin uusinta saatavilla olevaa versiota 2.0.0-beta5. Aineisto luettiin ensin MySQL-tietokannasta SIARD 2.0 -muotoiseksi tiedostoksi, joka kokeiltiin palauttaa sekä toisella nimellä samaan MySQL-kantaan että PostgreSQL-tietokantaan (versio 9.3.14).

SMEAR-tietokannan muuntaminen SIARD-muotoon onnistui ongelmitta, ja tiedosto näytti ainakin lyhyen silmämääräisen tarkastelun perusteella sisältävän kaikki olennaiset tiedot. Muuntamisessa takaisin MySQL- ja PostgreSQL-tietokantoihin havaittiin muutamia ongelmia, jotka tullaan selvittämään ohjelmiston kehittäjien kanssa. Itse tiedostomuoto soveltunee SMEAR-aineiston sekä muiden samankaltaisten aineistojen säilytykseen, kunhan muunnos-työkalujen virhetilanteiden käsittelyyn liittyvät puutteet korjataan.

Muut tietokannat

Relaatiotietokantojen lisäksi on olemassa myös muun tyyppisiä tietokantoja, joita kutsutaan usein yleisnimellä NoSQL-kannat. Ne perustuvat kaksiulotteisten taulujen sekä niiden välisten suhteiden sijaan johonkin muuhun tietomalliin, esimerkiksi avain-arvo-pareihin, dokumenttien tai objektien säilytykseen. Relaatiotietokantojen tapaan myös NoSQL-kantoja käytetään yleensä kyselykielellä, jonka avulla voidaan tallentaa, hakea ja ladata tietoja kannasta. Kielet eivät kuitenkaan toistaiseksi ole SQL-kielen tapaan standardoituja. Tunnettuja NoSQL-kantoja ovat mm. Google BigTable, Amazon Dynamo ja avoimen lähdekoodin MongoDB.

Muiden kuin relaatiotietokantojen käytöstä tutkimusaineistojen tallennuksessa on toistaiseksi niukasti tietoja, eikä kukaan haastatelluista maininnut käyttävänsä tai tuntevansa NoSQL-kantoja. Niiden säilytykseen ei tässä selvityksessä paneuduta tarkemmin. Asiaan on syytä palata, mikäli NoSQL-kantoihin tallennettuja, pitkäaikaissäilytyksen kannalta arvokkaita tutkimusaineistoja tulee vastaan.

Aineistojen säilytyksen lisäksi tietokantojen avulla voidaan toteuttaa www-pohjaisia hakupalveluja, joiden avulla laajoista aineistoista on helpompi löytää halutut osat. Itse aineisto on tallennettu perinteisesti tiedostoihin ja tietokanta toimii vain hakupalveluna. Pitkäaikaissäilytyksen kannalta on näissä tapauksissa arvioitava, onko itse tietokannassa arvokasta säilytettävää tietoa, vai riittääkö varsinaisen aineiston sisältävien tiedostojen säilytys.

7 AINEISTOJEN HYVÄKSYMINEN PITKÄAIKAISSÄILYTYKSEEN

Tässä luvussa esitetyt alustavat kriteerit ja hyväksymisvaatimukset on laadittu tutkimusaineistojen PAS-ratkaisua kehittävässä projektiryhmässä.

Tavoitteena on tehdä aineistojen toimittamisesta säilytykseen helppoa ja kätevää, jotta aineistot saadaan mahdollisimman laajasti ja nopeasti uudelleenkäytettäväksi. Samalla on kuitenkin varmistettava, että aineistot ovat muille tutkijoille käyttökelpoisia ja niiden yhteyteen on tallennettu kaikki säilytyksessä tarvittavat tiedot.

Lähtökohdaksi otettiin Kansallisen digitaalisen kirjaston pitkäaikaissäilytyksen (KDK-PAS) vaatimukset, joita muokattiin tutkimusaineistojen erityispiirteet huomioiden.

7.1 Säilytyksen tasot

Tutkimusaineistojen PAS-ratkaisussa on kaksi säilytystasoa:

1. Keskipitkä säilytys: aineiston eheyden säilytys ja julkaisu uudelleenkäytettäväksi
2. Pitkäaikaissäilytys: aineiston ymmärrettävyyden säilytys ja pitkäaikaisen saatavuuden varmistaminen

Pitkäaikaisella saatavuudella tarkoitetaan useita kymmeniä tulevia vuosia, joiden aikana mm. teknologia ja tutkimuskäytännöt muuttuvat.

Vaatimukset täyttävä aineisto voidaan siirtää suoraan pitkäaikaissäilytykseen, joka sisältää kaikki keskipitkän säilytyksen toiminnot. Vaihtoehtoisesti aineisto voidaan julkaista ensin keskipitkän säilytyksen palvelussa, jolloin sen siirtämisestä varsinaiseen pitkäaikaissäilytykseen päätetään myöhemmin.

Kun aineisto hyväksytään säilytykseen, PAS-ratkaisu luo sille pysyvän tunnisteen.

7.2 Vaatimukset aineiston hyväksymiseksi säilytykseen

Suurin osa vaatimuksista on samoja sekä keskipitkään että varsinaiseen pitkäaikaissäilytykseen vastaanotettaville aineistoille.

Keskipitkän säilytyksen osalta säilytykseen siirtoa on helpotettu lieventämällä tiedostomuotoihin liittyviä vaatimuksia varsinaiseen pitkäaikaissäilytykseen verrattuna. Aineisto kaikkine siihen kuuluvine osineen tulee kuitenkin myös keskipitkässä säilytyksessä olla kuvailtu asianmukaisesti.

Vaatimukset aineiston hyväksymiseksi säilytykseen on lueteltu alla.

1. Aineistokokonaisuus on muille tutkijoille käyttökelpoinen. (pakollinen)
   - Kokonaisuuteen tulee sisältyä kaikki ymmärrettävyyden kannalta olennaiset tiedot, mukaan lukien tutkimusmenetelmien ja tiedostojen dokumentaatio.
   - Kokonaisuuden tulee olla itsensä selittävä siten, että muut tutkijat voivat itsenäisesti hyödyntää aineistoa. Maallikolle sen ei tarvitse olla ymmärrettävä.
2. Kokonaisuuteen kuuluvat tiedostot ja niiden suhteet on kuvattu PAS-ratkaisun määritysten mukaisesti. (pakollinen)
   - Kuvaus tehdään osaksi METS-dokumenttia, ks. [KDK_Standardisalkku].
   - Haluttaessa METS-dokumentti voidaan laatia käyttämällä paketointipalvelua.
3. Tiedostot ovat joissakin PAS-ratkaisussa hyväksytyistä säilytys- tai siirtokelpoisista tiedostomuodoista. (pakollinen, keskipitkässä säilytyksessä suositeltava)
   - Mikäli jotkut tiedostoista eivät ole ennalta hyväksytyissä tiedostomuodoissa, niiden tulee täyttää erilliset tiedostomuotoja koskevat vaatimukset.
4. Aineiston käyttöoikeustiedot on ilmoitettu. (pakollinen)

- Tiedot ilmoitetaan kuvailupalvelun avulla, jossa on useimpiin tilanteisiin sopivat valmiit valinnat.

5. Aineiston lisenssi on voimassa olevien avoimen tieteen suositusten mukainen. (suositeltava)
    - Kuvailupalvelussa on lista suositeltavista lisensseistä. Tällä hetkellä voimassa oleva lisenssisuositus on Creative Commons Nimeä 4.0 (CC-BY 4.0).
6. Aineisto on kuvailtu metatietomääritysten mukaisesti. (pakollinen)
    - Kuvailuun on käytettävissä kuvailupalvelu.

Tarkemmat määritykset metatietojen ja käyttöoikeuksien kirjaamisen osalta laaditaan myöhemmin.

7.3 Tiedostomuotoja koskevat vaatimukset

Nämä vaatimukset koskevat niitä keskipitkään säilytykseen siirrettäviä tiedostoja, jotka eivät ole jossakin PAS-ratkaisussa hyväksytyistä säilytys- tai siirtokelpoisista tiedostomuodoista. Säilytys- ja siirtokelpoisten tiedostomuotojen osalta vaatimusten täyttyminen on jo varmistettu.

1. Tiedostomuoto on tuettu vähintään yhdessä yleisesti saatavilla olevassa ohjelmassa (pakollinen)
    - Ohjelma voi olla maksullinen tai maksuton. Mikäli tiedostojen käsittelyyn tarvitaan erikoisohjelma, tulee ohjelman nimi ja linkki ohjelman kotisivulle mainita.
    - Mikäli vain tutkijan itse kehittämä ohjelma tukee tiedostomuotoa, ohjelma tai sen lähdekoodi on liitettävä mukaan aineistokokonaisuuteen.
2. Tiedostomuodon rakenne on dokumentoitu (suositeltava)
    - Valmistajakohtaisia, suljettuja tiedostomuotoja voidaan siirtää keskipitkään säilytykseen, mutta niille ei voida taata ymmärrettävyyden säilymistä pitkällä aikavälillä
    - Mikäli mahdollista, aineisto tulee tallentaa suljettujen tiedostomuotojen rinnalle myös avoimessa, dokumentoidussa tiedostomuodossa.
    - Itse laaditut tiedostomuodot tulee dokumentoida dokumentointiohjeen mukaisesti.
3. Tiedostomuoto on alalla laajasti käytetty (suositeltava)
4. Tiedostomuoto on riippumattoman organisaation tai alan tutkimusyhteisön standardoima (suositeltava).

Vaatimusten mukaiset tiedostot voidaan siirtää ilman ennakkohyväksyntää keskipitkään säilytykseen. Palvelussa tiedostomuoto arvioidaan säilytys- ja siirtokelpoisten tiedostomuotojen valintaprosessissa, joka tekee päätöksen sen lisäämisestä hyväksyttyjen tiedostomuotojen listaan.

7.4 Säilytys- ja siirtokelpoisten tiedostomuotojen valintakriteerit

Säilytys- ja siirtokelpoisten tiedostomuotojen valinnassa käytetään seuraavia arviointikriteerejä:

1. Tiedostomuoto täyttää keskipitkään säilytykseen siirrettäviä tiedostomuotoja koskevat vaatimukset (tärkeä)
2. Tiedostomuoto on tuettu vähintään yhdessä avoimen lähdekoodin ohjelmassa (tärkeä)
3. Tiedostomuoto on laajasti tuettu eri ohjelmissa (melko tärkeä)

4. Tiedostomuodon dokumentaatio on selkeä ja laadukas (tärkeä)
5. Tiedostomuodon dokumentaatio on saatavilla maksutta (ei kovin tärkeä)
6. Tiedostomuoto on alas- ja ylöspäin yhteensopiva (ei kovin tärkeä)
7. Tiedostomuoto on kansainvälisesti valittu säilytyskelpoiseksi ainakin jossain tunnetussa säilytyspalvelussa / data-arkistossa (melko tärkeä)
8. Tiedostomuoto on vakaa, uusia versioita julkaistaan vain harvoin (ei kovin tärkeä).

Kriteerit pohjautuvat KDK-PAS-hankkeessa laadittuihin valintakriteereihin. KDK-PAS:n kriteerien soveltuvuutta tutkimusaineistoille on arvioitu tarkemmin liitteessä D.

7.5 Huomioita liittyen aineistojen hyväksymiseen säilytykseen

Vaatimuksissa mainitun METS-dokumentin muodostaminen voidaan tehdä ohjatusti osana aineiston paketointipalvelua. Vaihtoehtoisesti aineiston lähettäjä voi huolehtia asiasta omassa tietojärjestelmässään ja toimittaa palveluun valmiin METS-tiedoston.

Metatietomääritykset ovat osittain tiedostomuotokohtaisia. Määrityksiä kannattaa kuitenkin pyrkiä yhtenäistämään niin paljon kuin mahdollista, jotta eri aineistojen yhdistäminen ja käyttö yli tieteenalarajojen helpottuu. Esimerkiksi merkistöjen osalta voitaneen vaatia tai ainakin suositella lähes kaikilla aloilla de facto -standardiksi muodostuneen UTF-8-merkistön käyttöä.

Esitettyjen vaatimusten täyttyminen voidaan osittain varmistaa ohjelmallisesti validaattorien avulla. Aineistojen laadun varmistamiseksi saattaa kuitenkin olla tarpeen sisällyttää myös aineiston kuvauksen manuaalinen tarkistus osaksi vastaanotto- ja julkaisuprosessia. PAS-ratkaisun tulee tukea prosessia, jossa aineiston lähettäneen henkilön lisäksi toinen henkilö tarkastaa ja hyväksyy sen.

Joissain tapauksissa tiedostomuodolle voi olla useita vaihtoehtoja, jolloin PAS-ratkaisu voi ohjata käyttämään säilytyksen ja uudelleenkäytön kannalta parhaita mahdollisia muotoja. Esimerkiksi rakenteiset tekstitiedostot voidaan hyväksyä säilytykseen, mutta suositella niiden sijaan XML-pohjaisia tiedostomuotoja. Suositeltavien muotojen käyttöön voidaan ohjata laatimalla niille palvelussa laajennettu tuki — mahdollisuus valita kuvailu- tai paketointipalvelussa valmiista listasta tai automaattinen tunnistus tiedoston lähettämisen yhteydessä, jolloin tiedostomuodon dokumentaatio tulee automaattisesti mukaan.

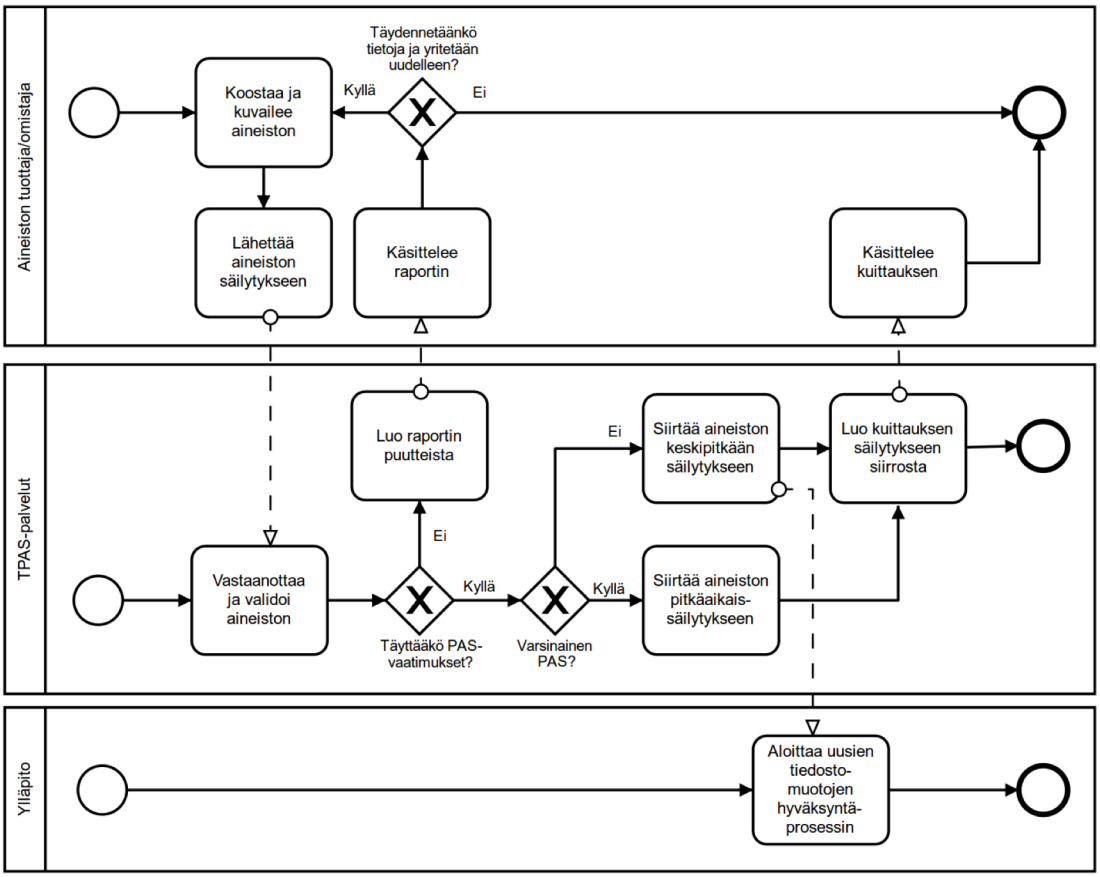
Osalle tutkimusaineistoista saattaa riittää keskipitkä säilytys. Elinkaaren pituutta on kuitenkin usein vaikea arvioida heti aineiston valmistuttua. Aineiston osoittautuessa suosituksi ja tiedostomuotojen kehittyessä voidaan päätös varsinaiseen pitkäaikaissäilytykseen siirtämisestä tehdä myöhemmin, tarvittaessa jopa useita vuosia sen jälkeen kun aineisto on siirretty keskipitkään säilytykseen ja julkaistu sen puitteissa uudelleenkäytettäväksi.

Keskipitkän säilytyksen kestosta on syytä tehdä selkeä päätös. Esimerkiksi Suomen Akatemiaa vastaava saksalainen Deutsche Forschungsgemeinschaft (DFG) on linjannut hyvän tutkimuskäytännön edellyttävän kymmenen vuoden säilytysaikaa [DFG_Praxis]. Asiallisesti dokumentoitujen aineistojen voitaneen arvioida säilyvän kymmenen vuotta käyttökelpoisina ilman tiedostomuotojen muunnoksia tai muita suurempia toimenpiteitä, kunhan eheydestä huolehditaan. Keskipitkän säilytyksen määräajan päättyminen ei välttämättä tarkoita tietojen poistamista, mutta aineiston käyttökelpoisuutta ei seurata eikä sille tehdä ymmärrettävyyden säilymisen vaatimia toimenpiteitä kuten varsinaisessa pitkäaikaissäilytyksessä.

Tutkimusaineistojen säilytykseen siirrosta vastaava organisaatio saattaa olla tutkimusinfrastruktuuri. Ne hallinnoivat tyypillisesti tietyn tieteenalan aineistoja yliopistorajat ylittäen, ja niillä on siten paremmat valmiudet dokumentoida niitä yhtenäisesti kuin yliopistoilla tai yksittäisillä tutkijoilla. Tutkimusinfrastruktuurien kanssa on myös tärkeää tehdä yhteistyötä valittaessa säilytys- ja siirtokelpoisia tiedostomuotoja. Toisaalta on huomioitava, että infrastruktuureilla on usein oma tallennuspalvelu, jonka rooli PAS-ratkaisuun nähden on syytä selvittää.

7.6 Hyväksymisprosessi

Hahmotelma aineistojen hyväksymisprosessiksi on esitetty kuvassa 3. Siinä näkyy myös keskipitkän säilytyksen rooli.



Kuva 3: Hahmotelma aineistojen hyväksymisprosessiksi

Aineiston tuottaja tai omistaja koostaa, kuvailee ja paketoi säilytykseen siirrettävän aineistokokonaisuuden joko omassa tietojärjestelmässään tai TPAS-kokonaisuuteen kuuluvia palveluja (kuvailu- ja paketointipalvelu, ei esitetty kuvassa) hyödyntäen. Sen jälkeen hän siirtää sen TPAS-palveluihin kuuluvaan PAS-järjestelmään, joka vastaanottaa ja validoi aineiston.

Vaatimukset täyttänyt, validoinnin läpäissyt aineisto siirretään joko keskipitkään tai varsinaiseen pitkäaikaissäilytykseen. Valinta niiden välillä voi riippua tehdyistä säilytyssopimuksista tai teknisten vaatimusten täyttymisestä. Ehdotetussa mallissa suurin ero on tiedostomuotojen ennakkohyväksyntä, jota edellytetään vain varsinaiseen pitkäaikaissäilytykseen siirrettäviltä aineistoilta. Keskipitkän säilytyksen tapauksessa ylläpito saa tiedon aineistossa esiintyvistä uusista tiedostomuodoista, ja käynnistää kyseisten tiedostomuotojen hyväksyntäprosessin. Sen tulos ratkaisee osaltaan, siirretäänkö aineisto myöhemmin keskipitkästä säilytyksestä varsinaiseen pitkäaikaissäilytykseen (ei esitetty kuvassa).

Aineiston säilytykseen siirtänyt taho saa sekä keskipitkän että varsinaisen pitkäaikaissäilytyksen tapauksessa kuittauksen siirron onnistumisesta. Mikäli aineisto ei täytä vaatimuksia tai validointi epäonnistuu, toimitetaan kuittauksen sijaan raportti puutteista. Jos vastaanotto, validointi tai siirto epäonnistuu käyttäjästä riippumattomista teknisistä syistä, ylläpito ryhtyy toimenpiteisiin. Sen tehtäviin kuuluu myös asiakaspalvelu ongelmatilanteissa. Niitä ei kuitenkaan ole esitetty kuvassa, jotta prosessin normaali kulku on selkeämmin nähtävissä.

Prosessiin saattaa olla tarpeen lisätä automaattisen validoinnin lisäksi ihmisen suorittama hyväksyntä, jossa esimerkiksi tarkastetaan kuvailun laatu. Yksityiskohtia tarkennetaan tutkimusaineistojen pitkäaikaissäilytyksen toimijoiden ja vastuualueiden selkiytymisen myötä.

Joka tapauksessa sekä hyväksyntäprosessin sekä PAS-ratkaisun tulee tukea vastuun jakoa eri toimijoille aineistojen tarkastamisen, tiedostomuotojen seurannan ja muunnosten osalta.

7.7 Esimerkkiaineistojen valmius pitkäaikaissäilytettäväksi

Tässä osiossa on arvioitu, kuinka valmiita selvityksessä mukana olleet esimerkkiaineistot ovat pitkäaikaissäilytyksen kannalta. Arviot valmiudesta pohjautuvat KDK-PAS:n määrityksiin sekä tämän selvityksen osana laadittuihin alustaviin tutkimusaineistojen säilytysvaatimuksiin. Taulukkoon on myös kirjattu tarvittavat muutokset, jotta aineistot voitaisiin vastaanottaa säilytettäväksi alustavien hyväksymisvaatimusten mukaisesti.

Kyseessä ei ole aineistojen paremmuusjärjestys, vaan esimerkkien avulla tehty katsaus siitä, millaisia muutostöitä on odotettavissa valmisteltaessa aineistoja säilytystä varten. Keskimäärin työtä lienee selvityksessä mukana olleita aineistoja enemmän, koska osa mukaan valituista aineistoista oli mukana Tutkimuksen PAS-palvelun piloteissa [PAS_Pilotit_2015] ja niitä oli siten jo täydennetty pitkäaikaissäilytystä silmällä pitäen.

Aineisto	Valmius aineistokokonaisuuden ja dokumentaation osalta	Valmius tiedostomuotojen osalta
1000Gen	▪ Ymmärrettävyyden kannalta olennainen dokumentaatio ei ole tiedostojen mukana, mutta se olisi ainakin olennaisilta osin kerättävissä 1000 Genomes -projektin www-sivuilta. ▪ Aineisto olisi paketoitava säilytystä varten PAS-määritysten mukaisesti.	▪ Tiedostomuodot eivät ole kansallisesti eivätkä kansainvälisesti hyväksyttyjä siirto- tai säilytyskelpoisiksi, mutta dokumentoituja ja alalla vakiintuneita. ▪ Tiedostomuodot täyttävät keskipitkän säilytyksen vaatimukset ja suositukset.
Aivokuvat	▪ Aineisto oli mukana PAS-pilotissa ja sisältää ymmärrettävyyden kannalta olennaisen dokumentaation sekä paketointiin kuuluvan METS-tiedoston.	▪ Tiedostomuodot eivät ole kansallisesti eivätkä kansainvälisesti hyväksyttyjä siirto- tai säilytyskelpoisiksi, mutta dokumentoituja ja alalla vakiintuneita. ▪ Tiedostomuodot täyttävät keskipitkän säilytyksen vaatimukset ja suositukset
ERNE	▪ Aineisto oli mukana PAS-pilotissa ja sisältää ymmärrettävyyden kannalta olennaisen dokumentaation sekä paketointiin kuuluvan METS-tiedoston.	▪ Data ei ole standardoidussa tai alalla vakiintuneessa tiedostomuodossa. Muoto on kuitenkin dokumentoitu. ▪ Dokumentit ja kuvat ovat säilytyskelpoisissa tiedostomuodoissa. ▪ Muut tiedostomuodot täyttävät keskipitkän säilytyksen vaatimukset.

Aineisto	Valmius aineistokokonaisuuden ja dokumentaation osalta	Valmius tiedostomuotojen osalta
FIRE	▪ Aineisto sisältää mittaus-kokonaisuuden ymmärtämisen kannalta olennaisen dokumentaation. ▪ Mittauksissa käytettyjen parametrien dokumentointia tulisi säilytystä varten vielä parantaa. ▪ Aineisto olisi paketoitava säilytystä varten PAS-määritysten mukaisesti.	▪ Datan tiedostomuoto SEG-Y ei ole kansallisesti eikä kansainvälisesti hyväksytty siirto- tai säilytyskelpoiseksi, mutta dokumentoitu ja alalla vakiintunut. ▪ SEG-Y:n lisäksi suurin osa muista tiedostomuodoista täyttää vähintään keskipitkän säilytyksen vaatimukset. ▪ Oheistiedostot pääosin rakenteisia tekstitiedostoja, jotka ovat KDK:ssa säilytyskelpoisia normaalina tekstinä. Tiedostojen rakenne syytä kuitenkin dokumentoida paremmin. ▪ Loppuraportti kannattaa säilytystä varten muuntaa kokonaisuudessaan säilytyskelpoiseen PDF-muotoon.
FSD	▪ Aineisto sisältää ymmärrettävyyden kannalta olennaisen dokumentaation. ▪ FSD:llä on oma yhtenäinen aineistojen kuvailu- ja paketointikäytäntö, jonka pohjalta on helppo luoda myös PAS-ratkaisun määritysten mukainen paketointi.	▪ Datan tiedostomuodot RTF ja SPSS Portable eivät toistaiseksi ole KDK:ssa siirto- tai säilytyskelpoisiksi hyväksyttyjä. ▪ SPSS Portable ollaan KDK:ssa alustavasti hyväksymässä lähiaikoina tietyin reunaehdoin. ▪ Kansainvälisesti RTF on laajasti hyväksytty, SPSS on hyväksytty joissakin organisaatioissa. ▪ Dokumentaatio ja muut oheistiedostot ovat kaikki valmiiksi säilytyskelpoisessa muodossa.
Kiteet	▪ Aineisto sisältää ymmärrettävyyden kannalta olennaisen dokumentaation. ▪ Aineisto olisi paketoitava säilytystä varten PAS-määritysten mukaisesti.	▪ Data (tulostiedosto) on KDK:ssa siirtokelpoiseksi hyväksytyssä muodossa. ▪ Dokumentaatio on säilytyskelpoiseksi hyväksytyssä muodossa.

Aineisto	Valmius aineistokokonaisuuden ja dokumentaation osalta	Valmius tiedostomuotojen osalta
MAXIV	▪ Ymmärrettävyyttä hankala arvioida, koska kyseessä on Nexus HDF5 -määrityksen esimerkki eikä varsinaisen tutkimusprojektin aineisto. ▪ Aineisto olisi paketoitava säilytystä varten PAS-määritysten mukaisesti.	▪ HDF5-tiedostomuotoa ei ole hyväksytty KDK:ssa säilytys- tai siirtokelpoiseksi, kansainvälisesti hyväksytty osassa organisaatioista. ▪ HDF5-muoto ei sellaisenaan takaa ymmärrettävyyttä, vaan datatyypit ja metatiedot on myös määriteltävä. ▪ Nexus-määritys tarkentaa hyvin yleiskäyttöistä HDF5-määritystä ja siten parantaa säilytettävyyttä.
Planck	▪ Ymmärrettävyyden kannalta olennaiset dokumentit olisi alan asiantuntijan koostettavissa Planck-arkiston sivuilta (esimerkkiaineiston valitsi sivuilta selvityksen tekijä). ▪ Aineisto olisi paketoitava säilytystä varten PAS-määritysten mukaisesti.	▪ Datan tiedostomuoto FITS ei ole kansallisesti eikä kansainvälisesti hyväksytty siirto- tai säilytyskelpoiseksi, mutta alalla vakiintunut, hyvin dokumentoitu ja riippumattoman työryhmän ylläpitämä. ▪ Muoto täyttää kaikki keskipitkän säilytyksen vaatimukset ja suositukset. ▪ Kuvat ovat säilytyskelpoiseksi hyväksytyssä PNG-muodossa.
RITU	▪ Aineisto oli mukana PAS-pilotissa ja sisältää ymmärrettävyyden kannalta olennaisen dokumentaation sekä paketointiin kuuluvan METS-tiedoston.	▪ Datan tiedostomuoto laitevalmistajan laatima, ei hyväksytty kansallisesti eikä kansainvälisesti siirto- tai säilytyskelpoiseksi. Muoto on kuitenkin dokumentoitu. ▪ Muut aineiston tiedostomuodot joko siirto- tai säilytyskelpoisia

Aineisto	Valmius aineistokokonaisuuden ja dokumentaation osalta	Valmius tiedostomuotojen osalta
SMEAR	▪ Ymmärrettävyyden kannalta olennaisten dokumenttien koostaminen vaatisi jonkin verran työtä (kaikki ei saatavilla samasta lähteestä). ▪ Aineisto olisi paketoitava säilytystä varten PAS-määritysten mukaisesti.	▪ Data on tallennettu MySQL-tietokantaan, josta se voidaan lukea MySQL-spesifiseen tiedostoon tai muuntaa paremmin säilytettäväksi soveltuvaan SIARD-muotoon. ▪ MySQL ei ole kansallisesti eikä kansainvälisesti siirto- tai säilytyskelpoiseksi hyväksytty, mutta dokumentoitu muoto. ▪ SIARD on säilytettävissä XML-tiedostona ja kansainvälisesti hyväksytty säilytyskelpoiseksi joissakin arkistoissa.
Suomi24	▪ Aineisto ei sisällä kaikkia ymmärrettävyyden kannalta olennaisia dokumentteja. Se on kuitenkin suurelta osin ymmärrettävissä tiedostoja tarkastelemalla.	▪ Datan tallennusmuoto VRT on rakenteinen tekstitiedosto, joka on KDK:ssa säilytettävissä normaalina tekstinä. Tiedostomuodon rakenteen dokumentointi on kuitenkin puutteellinen. ▪ Dokumentaatio on KDK:ssa säilytyskelpoisiksi hyväksyttyinä tekstitiedostoina.

Suurin osa esimerkkiaineistoista voitaisiin vastaanottaa ainakin keskipitkään säilytykseen varsin pienin muutoksin. Lähes kaikki tiedostomuodot täyttävät keskipitkän säilytyksen vaatimukset. Täydennettävää on lähinnä dokumentoinnissa, erityisesti rakenteisten tekstitiedostojen osalta. Tiedostomuotojen hyväksyminen siirto- tai säilytyskelpoisiksi vaatisi yksityiskohtaisempaa perehtymistä, sekä sen pohjalta laadittavia tarkempia määrityksiä erityisesti teknisten metatietojen osalta.

Aineistokokonaisuuksien kuvailut eivät ole keskenään vertailukelpoisia tai yhtenäisiä, koska kuvailun ohjeistus ja metatietomalli puuttuvat. Lisäksi dokumentaation laatua on vaikea arvioida ilman syvällistä tuntemusta kyseisistä tieteenaloista.

KDK-PAS:n määritysten mukainen paketointi METS-tiedostoineen puuttuu luonnollisesti kaikista muista paitsi TPAS-piloteissa mukana olleista esimerkkiaineistoista. Kaikki aineistot ovat kuitenkin paketoitavissa määritysten mukaisesti.

8 JOHTOPÄÄTÖKSET

KDK:n pitkäaikaissäilytyksen määritykset muodostavat hyvän pohjan tutkimusaineistojen pitkäaikaissäilytykselle. Olemassa olevia määrityksiä voidaan laajentaa kattamaan uudet aineistotyypit ja tiedostomuodot. Tutkimusaineistojen erityispiirteet edellyttävät kuitenkin muutamia suurempia muutoksia sekä vastuualueiden, prosessien että teknisten määritysten osalta.

8.1 Johtopäätökset tiedostomuodoista

Yleisesti tutkimusaineistojen tiedostomuodot ovat kehittymässä pitkäaikaissäilytyksen kannalta suotuisaan suuntaan. Lisääntyvän kansainvälisen yhteistyön ja tieteen avoimuuden myötä aineistoilla on yhä useammin muitakin käyttäjiä kuin sen alkuperäiset tuottajat. Tämä on monilla tieteenaloilla jo johtanut tiedostomuotojen parempaan dokumentointiin ja yhtenäistymiseen.

Tiedostomuotojen kirjo on kuitenkin kulttuuriaineistoja suurempi, ja monet muodoista ovat tiedealakohtaisia. Uusien tiedostomuotojen säilytyskelpoisuuden arviointi ja valinta tulee olla jatkuvaa, koska muodot kehittyvät tutkimusmenetelmien kehittymisen mukana.

KDK:ssa säilytyskelpoisiksi hyväksytyistä tiedostomuodoista poiketen tutkimusaineistojen tiedostomuodoille ei pääsääntöisesti ole valmiita teknisten metatietojen metatietoskeemoja. Niiden laatiminen vaatinee huomattavia resursseja, mutta panostus maksaa itsensä takaisin metatietojen yhtenäistymisen ja sen kautta parantuvan aineistojen uudelleenkäytettävyyden myötä.

Tutkimusaineistoissa esiintyy melko usein projekteissa itse laadittuja tiedostomuotoja, jotka tulee dokumentoida ennen aineistojen siirtoa säilytykseen. Tiedostot saattavat sinänsä olla esimerkiksi tekstimuotoisia ja siten säilytyskelpoisia, mutta niiden sisäinen rakenne on ymmärrettävyyden kannalta olennainen. Tällaisten tiedostomuotojen hyväksymiseksi säilytykseen on laadittava selkeät dokumentointiohjeet ja vaatimukset.

Tietokantamuotoisten aineistojen säilytystä ja uudelleenkäyttöä vaikeuttaa se, että tiedot pitää säilytystä varten lukea tietokannasta tiedostoon ja siirtää käyttöä varten taas takaisin, mikäli halutaan hyödyntää tietokannan monipuolisia mahdollisuuksia hakea ja valita osia aineistosta. Lisäksi suosittujen tietokantojen tuottamat tiedostomuodot ovat valmistajakohtaisia. Alun perin Sveitsin kansallisarkistossa kehitetty SIARD-muoto on tarjolla olevista vaihtoehdoista paras, ja se vaikuttaa olevan vakiintumassa tietokantojen säilytysmuodoksi. Helppoa ja kätevää tapaa tarjota säilytettyjä tietokantoja loppukäyttäjille uudelleenkäytettäväksi ei kuitenkaan toistaiseksi ole.

8.2 Johtopäätökset aineistojen vastaanottamisesta säilytykseen

KDK:n tiedostomuotojen ennakkohyväksyntää edellyttävä malli johtaisi helposti useiden aineistojen jäämiseen pitkäaikaissäilytyksen ulkopuolelle tiedostomuotojen hyväksyntäprosessin hitauden takia.

Säilytykseen siirtämistä voidaan helpottaa ottamalla käyttöön uusi keskipitkän säilytyksen taso, jossa tiedostomuotoihin liittyviä vaatimuksia on lievennetty KDK-PASiin verrattuna. Siten aineistot voidaan vastaanottaa säilytykseen, käynnistäen samalla rinnakkainen prosessi päättämään uusien tiedostomuotojen siirto- ja säilytyskelpoisuudesta sekä niihin liittyvistä metatietovaatimuksista. Näin aineistot saadaan nopeammin talteen ja uudelleenkäytettäväksi, ja päätös niiden siirtämisestä varsinaiseen pitkäaikaissäilytykseen voidaan tehdä myöhemmin.

Aineistojen ymmärrettävyyden osalta tutkimusaineistot poikkeavat kulttuuriaineistoista siten, että tulkinta vaatii usein syvällistä alan asiantuntemusta. Säilytykseen vastaanotettavien aineistojen ei siten tarvitse olla maallikolle ymmärrettäviä. Kuvailun ja dokumentaation tavoitteena ja vaatimustasona tulee olla se, että toinen tutkija ymmärtää ja voi hyödyntää aineistoa.

Osalle aineistoista saattaa riittää keskipitkä säilytys, jolloin varsinaisen pitkäaikaissäilytyksen resurssit voidaan kohdentaa suosituksi osoittautuneisiin tai muuten arvokkaiksi arvioituihin aineistoihin. Aineistojen kuvailun tulee kuitenkin jo keskipitkään säilytyksen siirrettäessä olla riittävä, jotta aineistokokonaisuus on muille tutkijoille käyttökelpoinen. Kuvailun ja muun dokumentaation täydentäminen jälkikäteen on hankalampaa kuin tiedostomuotoihin tehtävät tekniset tarkennukset ja muunnokset.

KDK:ssa laadittu aineistojen paketointimalli soveltuu hyvin myös tutkimusaineistoille. Kuvailu- ja paketointipalveluiden helppokäyttöisyyteen on syytä erityisesti panostaa. Tutkimusaineistojen erityishaasteita ovat eri tieteenalojen aineistojen huomattava poikkeaminen toisistaan, sekä joidenkin aineistojen erittäin suuri koko tai tiedostojen lukumäärä.

8.3 Johtopäätökset toimijoista ja vastuualueista

KDK:n aineistoista vastaa yleensä museo, kirjasto tai arkisto, jolla on lakisääteinen säilytystehtävä. Tutkimusaineistojen osalta tilanne ei ole yhtä selkeä. Aineistot on tyypillisesti laadittu tutkimushankkeissa, joilla on päättymispäivä eikä pitkäaikaista vastuuta aineistojen säilytyksestä. Lisäksi monet aineistot kerätään kansainvälisenä yhteistyönä, jolloin niitä ei omista mikään yksittäinen organisaatio. Joka tapauksessa aineistoja halutaan enenevässä määrin säilyttää ja julkaista, sitä edellyttävät myös yhä useammat tutkimuksen rahoittajat. Tutkimusaineistojen PAS-ratkaisulle on siten selkeä tarve.

Tutkimusaineistojen säilytykseen siirrosta vastaava organisaatio saattaa olla tutkimusinfrastruktuuri. Ne hallinnoivat tyypillisesti tietyn tieteenalan aineistoja yliopistorajat ylittäen, ja niillä on siten paremmat valmiudet dokumentoida aineistoja yhtenäisesti kuin yliopistoilla tai yksittäisillä tutkijoilla. Tutkimusinfrastruktuurien kanssa on myös syytä tehdä yhteistyötä valittaessa säilytys- ja siirtokelpoisia tiedostomuotoja. Toisaalta on huomioitava, että infrastruktuureilla on usein oma tallennuspalvelu, jonka rooli PAS-ratkaisuun nähden on syytä selvittää.

Tutkimusaineistojen pitkäaikaissäilytys on myös kansainvälisesti vielä melko varhaisessa vaiheessa. Joissain organisaatioissa on jo ehditty laatia kriteerejä tutkimusaineistojen säilytykseen ja hyväksyä joitakin tiedostomuotoja, mutta kattavaa listaa tarkkoine määrityksineen ei löydy yhdestäkään tarkastellusta organisaatiosta. Suurin osa myös keskittyy toistaiseksi aineistojen keskipitkään säilytykseen. Suomella on siten kansallisen PAS-ratkaisun myötä mahdollisuus olla edelläkävijä ja haluttu kumppani myös kansainvälisten tutkimusaineistojen säilytyksen osalta. Kansainvälinen yhteistyö on tutkimusaineistojen pitkäaikaissäilytyksessä KDK:takin tärkeämpää, koska yhä useammat aineistot tuotetaan kansainvälisesti ja niiden käyttö on maailmanlaajuista.

9 JATKOTYÖT

Tutkimusaineistojen pitkäaikaissäilytyksen suunnittelussa on vielä paljon työtä monella eri tasolla, korkean tason vastuualueiden määrittelystä palveluiden kehittämiseen ja erilaisiin teknisiin yksityiskohtiin. Alla on lueteltu erityisesti tutkimusaineistojen tiedostomuotoihin ja yleisemmin aineistojen säilytyskelpoisuuteen liittyviä tehtäviä. Lista ei ole tärkeysjärjestyksessä.

- Aineistokokonaisuuksien metatietomalli ja kuvailun ohjeistus. On tärkeää laatia tai valita yhteinen metatietomalli, jonka mukaisesti kaikkien aineistojen perustiedot voidaan kuvata. Kansainvälisen yhteistyön helpottamiseksi on syytä tukeutua valmiisiin malleihin, esimerkiksi EU:n suosittelemaan CERIF-malliin [CERIF].
- Tutkimusmenetelmien dokumentoinnin ohjeistus. Menetelmien kuvaukseen ei voida määritellä tiukkoja sääntöjä tai valmiita lomakkeita, mutta ohjeistuksen ja esimerkkien avulla voidaan helpottaa laadukkaiden ja ymmärrettävien kuvausten laatimista.
- Tiedostomuotojen hyväksyminen siirto- tai säilytyskelpoisiksi ja niihin liittyvät määritykset. Uusien tiedostomuotojen arviointi ja hyväksyminen on jatkuva prosessi, koska muodot muuttuvat ohjelmistojen ja tutkimusmenetelmien kehittymisen myötä. Työ on syytä aloittaa yleisistä ja vakiintuneista tiedostomuodoista, joihin tallennettuja aineistoja on jo valmiiksi tiedossa. Kullekin muodolle on KDK-PAS:n tapaan määriteltävä hyväksyttävät versiot, teknisten metatietojen skeema ja sen pakolliset sekä vapaaehtoiset kentät.
- Rakenteisten tekstitiedostojen ja itse kehitettyjen binäärimuotojen dokumentointiohjeet. Tiedostojen rakenne on olennainen aineistojen ymmärrettävyyden kannalta. Dokumentointiohjeiden avulla voidaan helpottaa aineistojen valmistelua säilytyskelpoisiksi sekä yhtenäistää käytäntöjä, mikä edistää uudelleenkäyttöä.
- Kuvailu- ja paketointipalvelun sekä tiedostomuotojen validoinnin kehittäminen. Palveluja on syytä kehittää ja pilotoida jo määritysten laatimisen rinnalla, jotta ne vastaavat käyttäjien tarpeita. Palveluiden käyttöönoton myötä määritysten sekä toimintamallien toimivuus tulee testattua käytännössä.
- Tutkimusaineistojen tiedostomuotojen systemaattinen kartoitus. Tässä selvityksessä asiaa tarkasteltiin esimerkkien kautta, mikä ei vielä kata kaikkia Suomessa tuotettujen tutkimusaineistojen tiedostomuotoja. Yksi tapa on maan laajuinen kartoitus kuten Itävallassa [Austrian_Survey], toinen vaihtoehto on perehtyä muotoihin tieteenalakohtaisesti lähestymällä useita kutakin tieteenalaa edustavia organisaatioita ja tutkimusryhmiä.

10 VIITTEET

[1000Genomes] The 1000 Genomes Project. http://www.1000genomes.org/

[AALTO] Lahnakoski, J. M., Salmi, J., Jääskeläinen, I. P., Lampinen, J., Glerean, E., Tikka, P., & Sams, M. (2012). Stimulus-Related Independent Component and Voxel-Wise Analysis of Human Brain Activity during Free Viewing of a Feature Film. PLoS ONE, 7(4), e35215. Public Library of Science. doi:10.1371/journal.pone.0035215 http://dx.doi.org/10.1371/journal.pone.0035215

[ATT_KVKatsaus] Tutkimusdatan pitkäaikaissäilytys: Kansanvälinen katsaus. Avoin tiede ja tutkimus -hanke, 2.11.2015. https://avointiede.fi/documents/10864/12232/Tutkimusdatan+pitk%C3%A4aikaiss%C3%A4ilytys+Kansainv%C3%A4linen+katsaus+2015.pdf

[Austrian_Survey] Bauer, Bruno; Ferus, Andreas; Gorraiz, Juan; Gründhammer, Veronika; Gumpenberger, Christian; Maly, Nikolaus; Mühlegger, Johannes Michael; Preza, José Luis; Sánchez Solís, Barbara; Schmidt, Nora; Steineder, Christian (2015): Researchers and their data. Results of an Austria survey – Report 2015. Version 1.2. DOI: 10.5281/zenodo.3400. https://phaidra.univie.ac.at/o:40931

[CDF_FAQ] CDF Frequently Asked Questions. http://cdf.gsfc.nasa.gov/html/FAQ.html

[CERIF] Common European Research Information Format (CERIF) - metatietomalli. http://www.eurocris.org/cerif/main-features-cerif

[CINES_Formats] CINES FACILE - Service de validation de formats, versio 3.4.4. https://facile.cines.fr/

[CLARIN_CMDI] CLARIN Component MetaData Infrastructure (CMDI). https://www.clarin.eu/content/component-metadata

[CoNLL-U] CoNLL-U-tiedostomuodon kuvaus. http://universaldependencies.org/format.html

[DANS_Formats] Data Archiving and Networked Services (DANS) Preferred Formats, September 2015, version 3.0. https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf

[DBPTK] Database Preservation Toolkit -ohjelmisto. http://www.database-preservation.com/

[DFG_Praxis] Deutsche Forschungsgesellschaft: Sicherung guter wissenschaftlicher Praxis / Safeguarding Good Scientific Practice, 25.10.2013. Print-ISBN 978-3-527-33703-3. DOI 10.1002/9783527679188.oth1. http://doi.org/10.1002/9783527679188.oth1

[DXFile] Data Exchange file format. http://dxfile.readthedocs.io/en/latest/

[ENA_SRA] European Nucleotide Archive (ENA): Sequence Read Archive (SRA) -metatietomallin kuvaus. http://www.ebi.ac.uk/ena/submit/metadata-model

[ERNE] ERNE-projektin ja mittalaitteen kuvaus. http://www.srl.utu.fi/projects/erne/

[FIRE_Hanke] FIRE-heijastusluotausprojekti. http://www.seismo.helsinki.fi/fi/tutkimus/proj/fire.html

[FSD2981] Kaikkien yhteinen kulttuuriperintö 2014 [elektroninen aineisto]. FSD2981, versio 1.0 (2015-01-15). Helsinki: Museovirasto & Helsinki: Suomen Kotiseutuliitto [tuottajat], 2014. Tampere: Yhteiskuntatieteellinen tietoarkisto [jakaja], 2015. https://services.fsd.uta.fi/catalogue/FSD2981

[FSD2985] Suomalaisten internetin käyttö 2013 [elektroninen aineisto]. FSD2985, versio 1.0 (2015-06-30). Helsinki: 15/30 Research & Helsinki: Yleisradio (YLE) [tuottajat], 2013. Tampere: Yhteiskuntatieteellinen tietoarkisto [jakaja], 2015. https://services.fsd.uta.fi/catalogue/FSD2985

[Human_Connectome] Human Connectome -projektin kotisivu. http://www.humanconnectome.org/

[KDK_Standardisalkku] Kansallisen digitaalisen kirjaston standardisalkku, 27.10.2014. http://www.kdk.fi/images/tiedostot/KDK_standardisalkku27_10_2014.pdf

[KDK_Tiedostomuodot] Kansallinen digitaalinen kirjasto: Säilytys- ja siirtokelpoiset tiedostomuodot, v1.4.0. http://www.kdk.fi/images/tiedostot/KDK-PAS-tiedostomuodot-v1.4.pdf

[LAC_Formats_2010] Library and Archives Canada (LAC), Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Access, Version 1.0, 2010. http://www.councilofnsarchives.ca/sites/default/files/LAC%20File%20Format%20Guidelines%20for%20Preservation%20and%20Long-term%20v1_2010-12_0.pdf

[LAC_Formats] Library and Archives Canada: Guidelines on File Formats for Transferring Information Resources of Enduring Value, 2015-02-05. http://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Documents/file-formats-irev.pdf

[LoC_Formats] Sustainability of Digital Formats, Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/

[LoC_Statement] Library of Congress Recommended Formats Statement 2016-2017. https://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf

[MAXIV_Lab] MAX IV -laboratorio. https://www.maxiv.lu.se/about-us/

[NAA_Formats] National Archives of Australia, Preservation file formats. http://naa.gov.au/Images/Preservation-File-Formats_tcm16-79398.pdf

[NetCDF] Network Common Data Form (NetCDF) -tiedostomuodon kotisivu. http://www.unidata.ucar.edu/software/netcdf/

[Nexus_HDF5_Talk] Nexus HD5 -tiedostomuodon kalvoesitys, Eugen Wintersberger, DESY, 27.05.2013. http://www.desy.de/dvsem/SS13/wintersberger_talk.pdf

[NNDC_Databases] National Nuclear Data Center Databases. http://www.nndc.bnl.gov/databases/databases.html

[ODA_DWG_Spec] Open Design Specification for .dwg files, version 5.3. http://opendesign.com/files/guestdownloads/OpenDesign_Specification_for_.dwg_files.pdf

[OGC_Standards] Lista Open Geospatial Consortium -järjestön standardeista. http://www.opengeospatial.org/standards

[PAS_Pilotit_2015] Tutkimuksen PAS-palvelun pilotit 2015: loppuraportti (8.4.2016) http://avointiede.fi/documents/10864/12232/Tutkimuksen+PAS-palvelun+pilotit+2015+Loppuraportti/3c0324c5-fb8c-4318-8582-c1600683dd78

[PLA] Planck Legacy Archive. http://pla.esac.esa.int/pla/

[PRONOM] PRONOM-tiedostomuotokirjasto. http://www.nationalarchives.gov.uk/PRONOM/

[RADWARE_FAQ] RadWare-analyysiohjelmiston usein kysyttyjen kysymysten lista (FAQ). http://radware.phy.ornl.gov/faq.html

[RITU] RITU gas filled recoil separator. https://www.jyu.fi/fysiikka/en/research/accelerator/nucspec/RITU

[ROOT] ROOT-analyysiohjelmiston kotisivu. https://root.cern.ch/

[SIARD_2004] Heuscher, Jaermann, Keller-Marxer, Moehle: Providing Authentic Long-term Archival Access to Complex Relational Data. Proceedings PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data, 5-7 October 2004, ESA/ESRIN, Frascati, Italy. Report number ESA WPP-232, pp. 241-261. http://arxiv.org/abs/cs/0408054

[SIARD_Standard] SIARD Format Specification v. 2.0, Sveitsin kansallinen standardi eCH-0165. http://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165

[SMEAR_AVAA] SMEAR-aineisto AVAA-palvelussa. http://avaa.tdata.fi/web/smart/smear

[Suomi24] Suomi 24 -korpus. http://urn.fi/urn:nbn:fi:lb-2015040801.

[UKDA_Formats] UK Data Archive File Formats Table. http://data-archive.ac.uk/create-manage/format/formats-table

LIITE A. HAASTATELLUT HENKILÖT

Projektissa haastatellut henkilöt on listattu alla olevassa taulukossa. Suurin osa haastatteluista tehtiin paikan päällä tai videoneuvotteluina. Kaksi haastatelluista vastasi kysymyksiin mieluummin sähköpostitse.

Organisaatio	Haastatellut henkilöt
Aalto-yliopisto, Neurotieteen ja lääketieteellisen tekniikan laitos, Aivo ja Mieli -laboratorio	Tutkija Enrico Glerean
Aalto-yliopisto, Kemian tekniikan korkeakoulu, Biotalousinfrastruktuuri	Varadekaani Sirkka-Liisa Jämsä-Jounela Lehtori Jukka Kortela
Biokeskus Suomi	Johtaja Olli Jänne Suunnittelija Marianna Jokila
Jyväskylän yliopisto, Fysiikan laitos, Kiihdytinlaboratorio	Tutkija Panu Rahkila
Helsingin yliopisto, Fysiikan laitos, Havaitsevan kosmologian ryhmä	Professori Hannu Kurki-Suonio Tutkija Elina Keihänen
Helsingin yliopisto, Fysiikan laitos, Ilmakehätieteet	Vastuullinen tutkija Ari Asmi Tutkijatohtori Pasi Kolari
Helsingin yliopisto, Geotieteiden ja maantieteen laitos, Seismologian instituutti	Tutkimusjohtaja Pekka Heikkinen Sovellussuunnittelija Kari Komminaho
Helsingin yliopisto, Nykykielten laitos	Tutkija Jussi Piitulainen
Lundin yliopisto (Ruotsi), MAX IV -laboratorio	IT-asiantuntija Krister Larsson
Oulun yliopisto, Fysiikan laitos, Nano- ja molekyylisysteemien tutkimusyksikkö	Professori Marko Huttula
Tieteen tietotekniikan keskus CSC	Paikkatietokoordinaattori Kylli Ek Sovellusasiantuntija Pekka Järveläinen Kehityspäällikkö Ilkka Lappalainen
Turun yliopisto, Fysiikan ja tähtitieteen laitos, Avaruustutkimuslaboratorio	Professori Rami Vainio
Yhteiskuntatieteellinen tietoarkisto FSD	IT-palveluasiantuntija Tuomas Alaterä Kehittämispäällikkö Mari Kleemola

LIITE B. HAASTATTELUKYSYMYKSET

Tässä liitteessä on esitetty haastatelluille henkilöille etukäteen lähetetyt kysymykset, jotka käytiin läpi haastattelun aikana. Mikäli henkilöillä ei ollut projektissa analysoitavaksi sopivaa esimerkkiaineistoa, keskityttiin kysymysten osioon B.

Kysymykset, osio A: Esimerkkiaineisto

Tässä osiossa käsitellään tutkimusryhmän omaa, esimerkiksi ajankohtaiseen tutkimusprojektiin liittyvää aineistokokonaisuutta. Aineisto voi koostua erilaisista tiedostoista, se voi olla tietokanta tai yhdistelmä edellisistä. Pyydämme teitä ehdottamaan sopivaa esimerkkiaineistoa, joka on mielestänne arvokas pitkäaikaissaatavuuden kannalta ja jota selvityksessä voidaan käsitellä.

Mikäli tietosuoja- ja tekijänoikeusrajoitusten puitteissa on mahdollista, aineistosta pyydetään toimittamaan otos, esimerkiksi kopio yhden kokeen tai mittaustapahtuman tiedostoista sekä niitä kuvailevat tiedot. Kopio voidaan tehdä haastattelun yhteydessä esimerkiksi USB-tikulle. Selvitystä tekevä työryhmä tarkastelee otoksen tiedostoja sekä arvioi niiden piirteitä pitkäaikaista tallennusta ja uudelleenkäyttöä ajatellen.

1. Aineistokokonaisuus

Mistä tiedostoista, tiedostomuodoista ja/tai tietokannoista aineisto koostuu?

Onko aineisto tallennettu tiettyyn hakemisto- tai muuhun rakenteeseen, jolla on merkitystä sen tulkinnan kannalta?

Kuinka suuri määrä aineistoa on kussakin muodossa?

2. Metatiedot

Mistä löytyvät aineiston tulkinnan kannalta olennaiset metatiedot (mm. tiedostojen rakenteen, mittalaitteiden asetukset, koetilanteen ym. kuvaavat tiedot)?

Ovatko kaikki mukana tiedostoissa vai onko olennaista tietoa esim. tutkimuksesta julkaistuissa artikkeleissa, erillisissä kuvailudokumenteissa tai tutkijoiden päässä (dokumentoimatonta hiljaista tietoa)?

3. Avoimuus, dokumentaatio ja standardit

Ovatko tiedostomuodot / rakenteet avoimia ja riittävästi dokumentoituja?

Onko tiedostomuodoille / rakenteille olemassa standardia?

Onko alalla olemassa standardeja, jotka on huomioitu tiedostomuotojen valinnassa, metatiedoissa, dokumentaatiossa tms.?

4. Käytetyt ohjelmat

Millä ohjelmilla aineistoa käsitellään?

Onko olemassa muita ohjelmia, joilla aineistoa pystyisi käsittelemään/tulkitsemaan?

5. Pysyvyys

Milloin joku aineiston tiedostomuodoista on viimeksi muuttunut?

Kuinka usein tiedostomuodot oman arvionne mukaan muuttuvat?

6. Yhteensopivuus

Onko tiedostoihin merkitty tiedostomuodon versionumero?

Onko uusin versio edellisen version / aiempien versioiden kanssa yhteensopiva (alaspäin/ylöspäin yhteensopiva)?

7. Korruptoitumisen sieto

Onko tiedostoissa tarkistussummia tai muuta mekanismia, jonka avulla mahdollinen korruptoituminen voidaan havaita?

Kuinka merkittävästi mahdollinen korruptoituminen vaikuttaa tiedoston tulkintaan?

8. Uudelleenkäyttö

Mihin aineisto on tällä hetkellä tallennettu?

Käytetäänkö kyseistä aineistoa joissain muissa tutkimusryhmissä?

Käytetäänkö samoja tiedostomuotoja joissain muissa tutkimusryhmissä / organisaatioissa?

Mitä seikkoja on syytä erityisesti huomioida, kun/jos joku toinen tutkija tai tutkimusryhmä käyttää/käyttäisi aineistoa?

Arvioitteko itse, että aineistolla saattaisi olla käyttäjiä 5, 10 tai 50 vuoden kuluttua?

9. Muut huomiot

Muuta huomioitavaa esimerkkiaineistoon liittyen?

Kysymykset, osio B: Tieteenalalla käytössä olevat tiedostomuodot

Nämä kysymykset koskevat joko kokonaista tieteenalaa (esim. fysiikka) tai tieteenalan alaosaa (esim. materiaalifysiikka, nanofysiikka), jolla haastateltava(t) henkilö(t) työskentelee / työskentelevät. Tässä siis kartoitetaan alalla käytössä olevia tiedostomuotoja, rakenteita ja tietokantoja laajemmin kuin osiossa A, jossa käsitellään ryhmän omaa esimerkkiaineistoa.

1. Yleisesti käytetyt tiedostomuodot

Mitä tiedostomuotoja alalla on yleisesti käytössä?

Osaatteko oman arvionne mukaan nimetä kaikki alan olennaisimmat tiedostomuodot, vai tiedättekö niistä vain osan?

Mistä muista lähteistä tai keiltä asiaa voisi selvittää tarkemmin?

2. Yleisesti käytetyt ohjelmistot

Mitä ohjelmistoja alalla on yleisesti käytössä?

Ovatko ne kaupallisten ohjelmistovalmistajien kehittämiä, alan tutkijoiden yhteistyössä kehittämiä vai yksittäisten tutkijoiden ja tutkimusryhmien kehittämiä ohjelmia?

3. Ohjelmistojen ja tiedostomuotojen avoimuus

Onko ohjelmistojen lähdekoodi (yleensä) saatavilla?

Ovatko rajapinnat hyvin dokumentoituja?

Ovatko tiedostomuodot hyvin dokumentoituja?

4. Yhteensopivuus

Ovatko tiedostomuodot alalla yhtenäisiä ja/tai yhteensopivia, vai onko heterogeenisuus ongelma?

Oletteko törmänneet tilanteeseen, jossa aineisto ei avaudu, esimerkiksi epäyhteensopivan tiedostomuodon tai tiedoston korruptoitumisen takia?

5. Metatiedot

Miten aineistojen tulkintaan liittyvät metatiedot (mm. tiedostojen rakenteen, mittalaitteiden asetukset, koetilanteen ym. kuvaavat tiedot) yleensä tallennetaan?

Mitä tietoja itse tarvitsette aineiston ymmärtämiseen, kun käytätte jonkun toisen tutkijan luomaa aineistoa?

6. Tietokantamuotoiset aineistot

Onko alalla käytössä tietokantamuotoisia aineistoja?

Onko kyseisiin tietokantoihin tarjolla hakuliittymiä, joiden kautta myös muut kuin tietokannan koonnut tutkijaryhmä/organisaatio pääsee käsiksi aineistoon?

Käytättekö itse tietokantamuotoisia aineistoja?

7. Standardit, määräykset ja ohjeet

Onko alalla ohjelmistoihin, tiedostomuotoihin tai metatietoihin liittyviä standardeja (virallisia tai de facto -standardeja)?

Antaako oma organisaationne ohjelmistoihin, tiedostomuotoihin tai metatietoihin liittyviä määräyksiä tai ohjeita?

Onko alalla joku muu auktoriteetti (esimerkiksi kansainvälinen organisaatio), joka antaa määräyksiä tai ohjeita?

8. Organisaatiot

Mitkä ovat merkittävimmät alan organisaatiot Suomessa / Euroopassa / maailmalla?

Onko teillä yhteistyötä heidän kanssaan ja/tai kontakteja kyseisissä organisaatioissa?

9. Aineistojen uudelleenkäyttö

Onko alan aineistoja jossain saatavilla uudelleenkäyttöä varten ja millaisin käyttöehdoin?

LIITE C. ESIMERKKIAINEISTOISSA ESIINTYVIEN TIEDOSTOMUOTOJEN ANALYYSI

Alla on kuvattu kaikki esimerkkiaineistoissa esiintyvät tiedostomuodot pitkäaikaissäilytyksen näkökulmasta. Huomiota on kiinnitetty erityisesti tiedostojen rakenteeseen, dokumentaation tasoon, mahdolliseen standardointiin, ohjelmistotukeen sekä luettavuuteen koneellisesti ja ihmissilmin.

Tiedot pohjautuvat haastatteluihin, esimerkkiaineistojen yksityiskohtaiseen tarkasteluun sekä verkkolähteisiin, erityisesti Library of Congressin tiedostomuotokirjastoon [LoC_Formats]. Tiedostojen vastaavuutta dokumentaation ja standardien kanssa tarkasteltiin vain silmämääräisesti, ilman koneellista validointia. Ohjelmistotuen osalta tukeuduttiin valmistajien ilmoituksiin ja muihin julkisesti saatavilla oleviin tietoihin, eikä ohjelmistoja kokeiltu muutamaa poikkeusta lukuun ottamatta. Kunkin tiedostomuodon osalta tarkastettiin, löytyykö muoto kansainvälisiltä suositeltavien tiedostomuotojen listoilta [CINES_Formats] [DANS_Formats] [LAC_Formats] [LoC_Statement] [NAA_Formats] [UKDA_Formats].

BAM / SAM

Koko nimi:	Binary Alignment/Map (BAM), Sequence Alignment/Map (SAM)
Uusin versio:	Versio 1 (18.11.2015) http://samtools.github.io/hts-specs/SAMv1.pdf
Avoimuus:	Avoin, dokumentoitu, ei-kaupallisen työryhmän kehittämä ja ylläpitämä
Yhteensopivuus:	Ei tietoa, toistaiseksi vain yksi versio
Ohjelmistotuki:	Useita eri ohjelmia, joilla tiedostoja voidaan käsitellä. Ks. esim. ELIXIR Tools and Data Services Registry, https://bio.tools/
Validointi:	Validaattoreita saatavilla. Eivät ilmeisesti toistaiseksi validoi kaikkia kenttiä, dokumentaatio puutteellinen. http://genome.sph.umich.edu/wiki/BamUtil:_validate http://broadinstitute.github.io/picard/command-line-overview.html#ValidateSamFile
Eheys:	Header-osiossa md5-tarkistussumma (vapaaehtoinen)
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	1000Gen-esimerkkiaineisto, geenisekvenssejä sisältävät aineistot
Huomioita:	▪ BAM on binäärinen, BGZF-pakattu versio SAM-tiedostoista. Muuten kyseessä on sama tiedostomuoto. ▪ Yksi alalla yleisesti käytetyistä tiedostomuodoista (muita mm. FastQ ja CRAM). ▪ BAM ei suoraan ihmisen luettavissa. Mikäli pakkaus puretaan esim. gzip-ohjelmalla, header-osion kentät luettavissa. ▪ Header-osiossa vain vähän standardin mukaan pakollisia kenttiä. Pitkäaikaissäilytyksen osalta määriteltävä, mitkä vapaaehtoiset kentät edellytetään täytettäväksi säilytettäviksi hyväksyttävissä tiedostoissa, sekä niiden sisältöön liittyvät yksityiskohdat. ▪ Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

BIDS

Koko nimi:	Brain Imaging Data Structure (BIDS)
Uusin versio:	1.0.0-rc2 http://bids.neuroimaging.io/bids_spec1.0.0-rc2.pdf
Avoimuus:	Avoin ja dokumentoitu, kansainvälisen työryhmän ylläpitämä
Yhteensopivuus:	Ei tietoa, toistaiseksi vain yksi versio
Ohjelmistotuki:	Ei toistaiseksi integroitu suurimpaan osaan ohjelmista. BIDS on rakenne eikä tiedostomuoto, joten käyttäjä tyypillisesti selaa rakennetta käyttöjärjestelmän normaaleilla työkaluilla kuten muitakin tiedostoja.
Validointi:	Validaattori saatavilla. https://github.com/INCF/bids-validator
Eheys:	Ei tarkistussummia. Validaattori tarkastaa rakenteen eheyden määritykseen verrattuna ja varoittaa poikkeamista tai puuttuvista arvoista.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	Aivokuvat-esimerkkiaineisto, MRI-kuvia sisältävät aineistot
Huomioita:	▪ BIDS ei ole tiedostomuoto vaan määritys, joka määrittää MRI-kuvia sisältävissä tutkimusaineistoissa käytettävän hakemistorakenteen, tiedostojen nimeämiskäytännöt, tiedostomuodot ja metatiedot. ▪ Alalla yleisesti käytetty ja hyväksytty rakenne, suunniteltu nimenomaan aineistojen uudelleenkäytön helpottamiseksi. ▪ Suppeahko joukko pakollisia tiedostoja ja metatietoja, huomattavasti laajempi joukko vapaaehtoisia (esim. erilaisia käytetyn kuvauslaitteiston tietoja ja parametreja). ▪ Luettavissa ja selattavissa sekä ihmissilmin että koneellisesti. ▪ Sallii myös määritysten ulkopuolisten tiedostojen sisällyttämisen hakemistorakenteeseen. ▪ Pitkäaikaissäilytyksessä käytettävä METS-rakennekartta on luultavasti mahdollista muodostaa suurelta osin automaattisesti, mikäli aineisto on BIDS-määrityksen mukainen. ▪ Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

CorelDraw (CDR)

Koko nimi:	CorelDraw
Uusin versio:	X8 / versio 18 (maaliskuu 2016)
Avoimuus:	Suljettu, ohjelmistokohtainen muoto, dokumentaatiota ei julkisesti saatavilla
Yhteensopivuus:	Alaspäin yhteensopiva
Ohjelmistotuki:	Vain CorelDraw-kuvankäsittelyohjelmisto tukee tiedostomuotoa täysin. Osittainen tuki avoimen lähdekoodin LibreOffice-ohjelmistossa.
Validointi:	Validaattoria ei saatavilla.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Kuvattu PRONOM-tiedostomuotokirjastossa, eri versiot erikseen http://www.nationalarchives.gov.uk/pronom/fmt/430 (versio X5)
Aineistot:	FIRE-esimerkkiaineisto
Huomioita:	▪ Suljettu, kaupallisen ohjelmiston vektorikuvatiedostomuoto, joka on pitkäaikaissäilytyksen ja uudelleenkäytön kannalta hankala. ▪ Muotoa lienee käytetty tutkimusaineistoissa lähinnä julkaisujen tai muiden dokumenttien kuvien laatimiseen ▪ CDR-muotoiset kuvat voidaan muuntaa esim. PDF- tai SVG-muotoon menettämättä kuvien tarkastelun kannalta olennaista informaatiota (mahdollisuus muokata kuvaa menetetään). ▪ DANSin suosituksissa mainitaan mahdollisuus avata CDR-tiedostot Adobe Illustrator -ohjelmalla ja muuntaa ne sitä käyttäen SVG-muotoon. ▪ Ei muita mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

CRAM

Koko nimi:	CRAM
Uusin versio:	Versio 3.0 (kesäkuu 2015) http://samtools.github.io/hts-specs/CRAMv3.pdf
Avoimuus:	Avoin, dokumentoitu, ei-kaupallisen työryhmän kehittämä ja ylläpitämä
Yhteensopivuus:	Alaspäin yhteensopiva vanhempien CRAM-tiedostojen sekä myös BAM-tiedostomuodon kanssa
Ohjelmistotuki:	Useita eri ohjelmia, joilla tiedostoja voidaan käsitellä. Ei kuitenkaan yhtä laaja ohjelmistotuki kuin BAM/SAM-muodolle.
Validointi:	Validaattoreita ei saatavilla.
Eheys:	Tarkistussummat käytössä.

LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	1000Gen-esimerkkiaineisto, geenisekvenssejä sisältävät aineistot
Huomioita:	▪ BAM / SAM -muodon pohjalta kehitetty tiedostomuoto, jonka päätavoitteena on tukea tehokkaampia pakkausmenetelmiä tilan säästämiseksi, tukien kaikkia BAM:n ominaisuuksia sekä tarjoten helpon siirtymäpolun BAM:sta CRAM:iin. ▪ Käytetään tyypillisesti häviöllisen pakkauksen kanssa, jossa geenisekvenssin informaatiosta jätetään hallitusti osia pois. ▪ Jonkin verran monimutkaisempi kuin BAM/SAM ▪ Yleistymässä ja tuettu useissa ohjelmakirjastoissa, ei kuitenkaan vielä yhtä laajasti kuin BAM/SAM. ▪ Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

DOC / DOCX

Koko nimi:	Microsoft Word Document (DOC), Office Open XML Document (DOCX)
Uusin versio:	ISO/IEC DIS 29500 (2012)
Avoimuus:	DOC suljettu, DOCX dokumentoitu ja standardoitu tiedostomuoto
Yhteensopivuus:	Alaspäin yhteensopiva
Ohjelmistotuki:	Tuettu useissa eri ohjelmissa. Kaikkien ominaisuuksien täysin toimiva tuki vain Microsoft Wordissa.
Validointi:	Validaattoreita ei saatavilla.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000397.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/412
Aineistot:	FIRE-esimerkkiaineisto, todennäköisesti yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).
Huomioita:	▪ Microsoft Word -tekstinkäsittelyohjelman käyttämä tiedostomuoto, joka on vähintään osittain tuettu myös monissa muissa ohjelmissa. ▪ Hyväksytty KDK:ssa siirtokelpoiseksi tiedostomuodoksi alkaen Word-ohjelmiston versiosta 97 (8.0). [KDK_Tiedostomuodot]. ▪ Kansainvälisesti laajasti hyväksytty siirtokelpoiseksi muodoksi (DANS, LAC, LoC, NAA, UKDA).

FITS

Koko nimi:	Flexible Image Transport System (FITS)
Uusin versio:	3.0 (heinäkuu 2008) http://fits.gsfc.nasa.gov/standard30/fits_standard30aa.pdf
Avoimuus:	Avoin, hyvin dokumentoitu, riippumattoman työryhmän ylläpitämä ja alan merkittävimpien organisaatioiden (mm. NASA, ESA) käyttämä.
Yhteensopivuus:	Alaspäin yhteensopiva
Ohjelmistotuki:	Ohjelmistokirjastoja saatavilla useille eri ohjelmointikielille
Validointi:	Validaattori saatavilla (FITSVerify) http://fits.gsfc.nasa.gov/fits_verify.html
Eheys:	Mahdollista lisätä tarkistussumma header-osioon. Rekisteröity käytäntö, ei kuitenkaan osa FITS-standardia. http://fits.gsfc.nasa.gov/registry/checksum.html
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/x-fmt/383
Aineistot:	Planck-esimerkkiaineisto, astronomista dataa sisältävät tutkimusaineistot
Huomioita:	▪ Jo 30 vuotta sitten kehitetty tiedostomuoto, joka on edelleen yleisesti käytössä astronomisen tiedon tallennuksessa. ▪ Melko monimutkainen rakenne, joka mahdollistaa kuvien lisäksi hyvin monenlaisen datan tallentamisen. ▪ Header-osio rakenteista tekstiä ja myös ihmisen luettavissa, varsinainen data binääristä. ▪ Header- ja data-osioita voi olla samassa tiedostossa myös useampia. ▪ Pitkäaikaissäilytyksen vaatimat tekniset metatiedot mahdollista tallentaa header-osioon. Määriteltävä pakolliset ja vapaaehtoiset kentät sekä niiden sisältöön liittyvät yksityiskohdat. Päätettävä, miten toimitaan useita header-osioita sisältävien tiedostojen kanssa, sekä mitkä laajennukset ovat tuettuja/sallittuja. ▪ Tulevan Euclid-satelliitin tuottaman datan tallentamisessa harkitaan siirtymistä FITS-muodosta HDF5-muotoon, lähinnä HDF5:n tarjoaman tehokkaamman pakkauksen takia. ▪ Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

GREAT

Koko nimi:	The GREAT / TDR Data Format
Uusin versio:	3.2.2 (lokakuu 2014) http://npg.dl.ac.uk/documents/edoc504/edoc504.html
Avoimuus:	Tiedostomuoto on dokumentoitu, mutta sen kehitys ei ole avointa. Laitevalmistaja julkaisee uusia versioita tai tarkennuksia tarpeen mukaan.
Yhteensopivuus:	Alaspäin yhteensopiva.
Ohjelmistotuki:	Kiihdytinlaboratoriossa itse kehitetty GRAIN-ohjelmisto (lähdekoodi saatavilla).
Validointi:	Laboratoriossa itse kehitetty validaattori.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	RITU-esimerkkiaineisto
Huomioita:	▪ Kyseessä on tutkimuksessa käytetyn GREAT-spektrometrin valmistajan kehittämä binäärimuoto, jonka dokumentaatio on saatavilla valmistajan kotisivuilta. ▪ Tiedostomuodossa itsessään ei ole paikkaa metatietojen tallentamiseen. Pitkäaikaissäilytyksen osalta kiinnitettävä huomiota siihen, että kaikki ymmärrettävyyden kannalta olennaiset metatiedot ja dokumentit ovat mukana aineistokokonaisuudessa. Tämä on pääosin jo tehty 2015 PAS-pilotin yhteydessä. ▪ Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

HDF5

Koko nimi:	Hierarchical Data Format 5 (HDF5)
Uusin versio:	HDF5 1.10, Specifications-dokumentin versio 2.0 https://www.hdfgroup.org/HDF5/doc/H5.format.html
Avoimuus:	Avoin ja dokumentoitu, ei-kaupallisen organisaation (The HDF Group) ylläpitämä
Yhteensopivuus:	Enimmäkseen sekä alas- ja ylöspäin yhteensopiva HDF5:n eri versioiden välillä, tietyt laajennukset eivät yhteensopivia. Edellinen pääversio HDF4 on täysin erilainen eikä HDF5 ole yhteensopiva sen kanssa. Konversiotyökalut HDF4->HDF5 ja HDF5->HDF4 on saatavilla.
Ohjelmistotuki:	Tuettu useissa eri ohjelmistoissa, joista läheskään kaikki eivät kuitenkaan tue kaikkia HDF5:n ominaisuuksia. Suurin osa käyttää tiedostojen lukemiseen HDF Groupin kehittämää avoimen lähdekoodin C-kirjastoa.

Validointi:	Validaattori saatavilla (HDF Group)
Eheys:	Mahdollisuus käyttää tarkistussummia, ei pakollinen ominaisuus. Korruption sietokyky yleisesti huono, pienikin korruptoituminen voi tehdä koko HDF5-tiedostosta käyttökelvottoman.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000229.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/807
Aineistot:	MAXIV- ja SMEAR-esimerkkiaineistot, yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).

Huomioita:

- HDF5 on yleiskäyttöinen tiedostomuoto, joka mahdollistaa lähes minkä tahansa tyyppisen datan tallentamisen.
- HDF5-tiedostoon voidaan tallentaa kahden tyyppisiä peruselementtejä: moniulotteisia taulukoita ja ryhmiä, joista kumpaankin voi lisäksi liittää attribuutteja. Näitä peruselementtejä hyödyntämällä voidaan tallentaa mm. kuvia, vektoreita, verkkoja ja metatietoja, sekä järjestää objektit halutulla tavalla puumaiseen rakenteeseen.
- Yleiskäyttöisyyden kääntöpuoli on monimutkaisuus, standardi on pitkä ja sen tukeminen kokonaisuudessaan vaativaa. Lisäksi on erilaisia laajennuksia ja lisämäärityksiä koskien mm. kuvien tallennusta.
- HDF5:n päälle on luotu eri hankkeissa tarkentavia määrityksiä, joissa kuvataan juuri kyseisen aineiston käyttämät datatyypit. Tällainen on mm. MAXIV-esimerkkiaineistossa käytetty Nexus HDF5.
- Pitkäaikaissäilytyksen kannalta on huomioitava, että HDF5-muodon käyttö ei sellaisenaan takaa ymmärrettävyyttä, vaan datatyypit ja metatiedot on myös määriteltävä, sekä tarkistettava PAS-järjestelmän vastaanotossa.
- Kansainvälisesti hyväksytty siirto- tai säilytyskelpoiseksi osassa tarkastelluista organisaatioista (CINES, DANS, LoC).

HTML

Koko nimi:	HyperText Markup Language (HTML)
Uusin versio:	HTML 5.0 (standardi) / HTML 5.1 (luonnos) https://www.w3.org/TR/html5/ https://www.w3.org/TR/html51/
Avoimuus:	Avoin, dokumentoitu, ei-kaupallisen organisaation (W3C) ylläpitämä
Yhteensopivuus:	Alaspäin yhteensopiva, pääosin myös ylöspäin yhteensopiva
Ohjelmistotuki:	Laajasti tuettu eri ohjelmissa, ei kuitenkaan aina kaikkien ominaisuuksien osalta. HTML-dokumentin visuaalisessa esittämisessä ohjelmakohtaisia eroja.
Validointi:	Validaattoreita saatavilla
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa

PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/471
Aineistot:	MAXIV-, Planck- ja SMEAR-esimerkkiaineistot, muut aineistot joissa ymmärrettävyyden kannalta olennaista dokumentaatiota esitetty www-sivuilla (tiedealariippumaton muoto).
Huomioita:	▪ HTML on www-sivujen laatimiseen ja esittämiseen käytetty merkintäkieli. ▪ HTML-muotoiset tiedostot sisältävät tyypillisesti linkkejä moniin muihin tiedostoihin (kuviin, äänitallenteisiin, ohjelmakoodiin ym.), jotka voivat olla missä tahansa tiedostomuodossa. Ilman näitä linkitettyjä tiedostoja kyseessä ei ole ymmärrettävä kokonaisuus. ▪ HTML-tiedosto itsessään voi sisältää Javascript-kielistä ohjelmakoodia, josta on oma määrityksensä ▪ Tutkimusaineistoissa HTML-muotoisia tiedostoja käytetään lähinnä dokumentaation tallentamiseen. ▪ KDK:ssa on hyväksytty säilytyskelpoiseksi tiedostomuodoksi sekä HTML (versio 4.01) että Web ARChive Format (WARC), joka kokoaa yhteen HTML-tiedostot sekä niihin liittyvät linkitetyt tiedostot [KDK_Tiedostomuodot]. ▪ Tutkimusaineistojen säilytyksen kannalta monissa tapauksissa mielekäs vaihtoehto on HTML-muotoisen dokumentaation muuntaminen PDF/A-muotoon, joka niin ikään on KDK:ssa säilytyskelpoiseksi hyväksytty tiedostomuoto. ▪ Kansainvälisesti laajasti hyväksytty säilytys- tai siirtokelpoiseksi muodoksi (DANS, LoC, NAA, UKDA).

Java

Koko nimi:	Java-ohjelmointikielen lähdekooditiedosto
Uusin versio:	Java SE 8
Avoimuus:	Avoin ja dokumentoitu. Java-kielen kehitystä kontrolloi Oracle-yritys, yhteisöllä mahdollisuus osallistua rajoitetusti.
Yhteensopivuus:	Alaspäin yhteensopiva
Ohjelmistotuki:	Useita eri Java-implementaatioita.
Validointi:	Validaattoria ei saatavilla.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	http://www.nationalarchives.gov.uk/pronom/x-fmt/422
Aineistot:	RITU-esimerkkiaineisto, muut aineistot joissa datan käsittelyyn käytetään Java-ohjelmointikieltä (tiedealariippumaton muoto)

Huomioita:	▪ Tekstimuotoinen tiedosto, joka sisältää Java-ohjelmointikielistä lähdekoodia. ▪ Tiedosto voidaan säilyttää tekstitiedostona (hyväksytty KDK:ssa säilytyskelpoiseksi tiedostomuodoksi), jolloin Java-kieltä hallitseva käyttäjä pystyy tulkitsemaan sitä. ▪ Tiedoston sisältämän lähdekoodin kääntäminen suoritettavaksi ohjelmaksi saattaa vaatia tietyn Java-version, jonka pitkäaikaissäilyttäminen on haastavaa. ▪ Kansainvälisesti niin ikään säilytettävissä tekstinä, ei mainintoja erityisestä Java-tuesta.

JPEG

Koko nimi:	Joint Photographic Experts Group (JPEG)
Uusin versio:	Versio 1.02 (syyskuu 1992) https://www.w3.org/Graphics/JPEG/jfif3.pdf
Avoimuus:	Avoin ja dokumentoitu, ISO-standardi
Yhteensopivuus:	Ei useampia versioita
Ohjelmistotuki:	Laaja tuki eri ohjelmistoissa
Validointi:	Validaattori saatavilla (jpeginfo) https://github.com/tjko/jpeginfo
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000018.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/44
Aineistot:	FIRE-esimerkkiaineisto, todennäköisesti yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto)
Huomioita:	▪ Häviöllistä pakkausta käyttävä kuvatiedostomuoto, hyväksytty KDK:ssa säilytyskelpoiseksi tiedostomuodoksi [KDK_Tiedostomuodot]. ▪ Kansainvälisesti laajasti hyväksytty säilytys- tai siirtokelpoiseksi muodoksi (CINES, DANS, LAC, LoC, NAA, UKDA).

JSON

Koko nimi:	JavaScript Object Notation (JSON)
Uusin versio:	Versio 1.0 https://tools.ietf.org/html/rfc7159
Avoimuus:	Avoin, dokumentoitu, IETF-standardi
Yhteensopivuus:	Ei useampia versioita
Ohjelmistotuki:	Laaja tuki erityisesti web-ohjelmistoissa
Validointi:	Useampia validaattoreita saatavilla
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.

LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000381.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/817
Aineistot:	Aivokuvat- ja SMEAR-esimerkkiaineistot, yhä yleisemmin käytössä myös muissa tutkimusaineistoissa (tiedealariippumaton muoto).
Huomioita:	▪ JSON on standardoitu, rakenteinen tekstitiedostomuoto, joka soveltuu erityisesti erilaisten metatietojen kuten mittausparametrien tallentamiseen sekä tietojen vaihtoon eri ohjelmien välillä. ▪ Muoto on alun perin kehitetty JavaScript-ohjelmointikielen osaksi, mutta nykyisin se on tuettu myös monissa muissa ohjelmointikielissä ja valmiissa ohjelmointikirjastoissa. ▪ Luettavissa sekä ihmissilmin että koneellisesti. ▪ Itse JSON-standardi määrittelee vain syntaksin, lisäksi on määriteltävä aineistokohtaisesti, mitä kenttiä ja arvoja JSON-tiedostoon tallennetaan. Tähän voisi soveltua JSON Schema (http://json-schema.org/). ▪ Voidaan käyttää pohjana johdannaisille tiedostomuodoille, joissa on määritelty esimerkiksi pakollisia kenttiä ja tarkennuksia tallennettavien arvojen syntaksiin (esim. GeoJSON). ▪ Säilytettävissä ainakin tekstitiedostona. Lisäksi mainittu erikseen LoC:n suosittelemissa muodoissa, osajoukko JSON-LD myös DANSin suosituksissa.

MySQL dump

Koko nimi:	MySQL dump file
Uusin versio:	5.7.15
Avoimuus:	Avoin ja dokumentoitu. Itse tiedostomuodosta ei ole erillistä dokumenttia, mutta se perustuu dokumentoituihin komentoihin, joiden avulla MySQL-tietokantaan syötetään tietoa. Kehittyy tietokannan kehityksen myötä kantaa kehittävän yrityksen (nykyisin Oracle) kontrolloimana.
Yhteensopivuus:	Alaspäin yhteensopiva.
Ohjelmistotuki:	MySQL-tietokannan mukana tuleva avoimen lähdekoodin mysqldump-ohjelma
Validointi:	Validaattoria ei saatavilla.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	SMEAR-aineisto
Huomioita:	▪ MySQL-tietokantojen varmuuskopiointiin tarkoitettu tiedostomuoto. ▪ Sisältää lyhyen rakenteellista tekstiä olevan header-osion.

- Loppuosa tiedostosta on lista SQL-komentoja, joiden avulla voidaan palauttaa alkuperäisen tietokannan taulut ja tiedot tyhjään kantaan. Komennot on MySQL:n ohjeissa hyvin dokumentoitu.
- Vanhemmasta kannasta luodut dump-tiedostot voidaan ainakin yleensä palauttaa kannan uudempaan versioon (alaspäin yhteensopiva).
- Muoto on MySQL-kohtainen eikä toimi muiden valmistajien kannoissa. Mysqldump-työkalun compatible-valitsimella voidaan tuottaa osittain yhteensopivia dump-tiedostoja, jotka eivät kuitenkaan yleensä toimi suoraan muissa tietokannoissa ilman käsin tehtäviä lisämuutoksia.
- Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

NIfTI

Koko nimi:	Neuroimaging Informatics Technology Initiative (NIfTI)
Uusin versio:	NIfTI 1.1 (2007) http://nifti.nimh.nih.gov/nifti-1 NIfTI 2.0: https://www.nitrc.org/docman/view.php/26/1302/Approved%20NIfTI-2%20Format%20document
Avoimuus:	Avoin ja dokumentoitu, kansainvälisen työryhmän ylläpitämä
Yhteensopivuus:	NIfTI 1.1 on sekä alas- että ylöspäin yhteensopiva version 1.0 kanssa NIfTI 2.0 ei ole yhteensopiva version 1.1 kanssa.
Ohjelmistotuki:	Useita muotoa tukevia ohjelmistoja saatavilla, sekä avoimen että suljetun lähdekoodin ohjelmistoja.
Validointi:	Validaattori saatavilla. https://github.com/INCF/bids-validator
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	Aivokuvat-esimerkkiaineisto, MRI-kuvia sisältävät aineistot

Huomioita:	▪ MRI-kuvatiedostojen, erityisesti aivoista otettujen MRI-kuvasarjojen tallentamiseen kehitetty muoto. Luodaan tyypillisesti muuntamalla MRI-skannerin tuottamat DICOM-tiedostot NIfTI-muotoon automaattisella muuntotyökalulla. ▪ BIDS-hakemistorakennemääritys edellyttää NIfTI-muodon käyttöä (joko versio 1.0/1.1 tai 2.0). ▪ Koneluettava muoto, ei tulkittavissa ihmissilmin. ▪ Header-osiossa vain vähän standardin mukaan pakollisia kenttiä. Pitkäaikaissäilytyksen osalta määriteltävä, mitkä vapaaehtoiset kentät edellytetään täytettäväksi säilytettäviksi hyväksyttävissä tiedostoissa, sekä niiden sisältöön liittyvät yksityiskohdat. ▪ Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

PDF

Koko nimi:	Portable Document Format (PDF)
Uusin versio:	PDF 1.7 (heinäkuu 2008) https://wwwimages2.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/PDF32000_2008.pdf PDF/A-3 (lokakuu 2012) http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229
Avoimuus:	Avoin ja dokumentoitu, kehitys pääosin Adobe-yhtiön kontrolloimaa
Yhteensopivuus:	Alaspäin yhteensopiva
Ohjelmistotuki:	Tuettu useissa eri ohjelmissa
Validointi:	Validaattoreita saatavilla
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000277.shtml (PDF 1.7) http://www.digitalpreservation.gov/formats/fdd/fdd000360.shtml (PDF/A-3)
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/276 (PDF 1.7) http://www.nationalarchives.gov.uk/pronom/fmt/479 (PDF/A-3a)
Aineistot:	Aivokuvat-, ERNE-, FIRE-, FSD-, MAXIV-, Planck-, ja RITU-esimerkkiaineistot, hyvin yleisesti käytössä myös muissa tutkimusaineistoissa (tiedealariippumaton muoto).

Huomioita:	▪ PDF-muodosta on perusversion lisäksi olemassa erikseen pitkäaikaissäilytystä varten kehitetty PDF/A-muoto, joka ei sisällä kaikkia perusversion ominaisuuksia. PDF/A-2 ja PDF/A-3 perustuvat PDF:n versioon 1.7. ▪ PDF/A:n versiot 1 ja 2 on KDK:ssa hyväksytty säilytyskelpoisiksi, ja PDF:n versiot 1.2-1.7 siirtokelpoisiksi tiedostomuodoiksi [KDK_Tiedostomuodot]. Siten suurin osa tutkimusaineistoihin kuuluvista PDF-tiedostoista lienee siirrettävissä pitkäaikaissäilytykseen ilman muutoksia. ▪ PDF/A:n versio 3:n olennaisin muutos versioon 2 verrattuna on tuki upotetuille (embedded) tiedostoille. Upotettujen tiedostojen säilytyskelpoisuutta standardi ei kuitenkaan määrittele. ▪ PDF/A-2- ja PDF/A-3-standardeista on kummastakin alaversiot a, b ja u, jotka asettavat eri tasoisia vaatimuksia dokumentin rakenteelle. Kaikki kolme PDF/A-2:n alaversiota ovat KDK:ssa hyväksyttyjä. ▪ Kansainvälisesti laajasti hyväksytty säilytyskelpoiseksi muodoksi (CINES, DANS, LAC, LoC, NAA, UKDA), versiotuki vaihtelee.

PNG

Koko nimi:	Portable Network Graphics (PNG)
Uusin versio:	ISO/IEC 15948:2003 (marraskuu 2003), vastaa olennaisilta osin versiota 1.2 https://www.w3.org/TR/PNG/
Avoimuus:	Avoin, dokumentoitu, ISO-standardi (ISO/IEC 15948:2003)
Yhteensopivuus:	Vähintään alaspäin yhteensopiva
Ohjelmistotuki:	Laajasti tuettu monissa eri ohjelmissa ja ohjelmointikirjastoissa
Validointi:	Validaattoreita saatavilla, mm. pngcheck http://www.libpng.org/pub/png/apps/pngcheck.html
Eheys:	CRC-32-tarkistussummat käytössä
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000153.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/13
Aineistot:	ERNE-esimerkkiaineisto, todennäköisesti yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).
Huomioita:	▪ Häviötöntä pakkausta käyttävä kuvatiedostomuoto. ▪ Hyväksytty KDK:ssa säilytyskelpoiseksi tiedostomuodoksi [KDK_Tiedostomuodot]. ▪ Kansainvälisesti laajasti hyväksytty säilytyskelpoiseksi muodoksi (CINES, DANS, LAC, LoC, NAA).

RTF

Koko nimi:	Rich Text Format (RTF)
Uusin versio:	1.9.1 (maaliskuu 2008)
Avoimuus:	Avoin ja dokumentoitu, kehitys yhden yrityksen (Microsoft) kontrolloimaa
Yhteensopivuus:	Alaspäin yhteensopiva
Ohjelmistotuki:	Varsin laajasti tuettu tekstinkäsittelyohjelmistoissa
Validointi:	Validaattoreita ei saatavilla
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/355
Aineistot:	FSD-esimerkkiaineisto, todennäköisesti käytössä myös muissa, erityisesti vanhemmissa tutkimusaineistoissa.
Huomioita:	▪ Pitkään melko laajasti käytössä ollut tiedostomuoto muotoiluja ja kuvia sisältävän tekstin tallennukseen. ▪ Koneluettava, jossain määrin myös ihmissilmin luettavissa. ▪ Periaatteessa tuettu useissa eri ohjelmissa ja siten soveltuu hyvin tietojen vaihtoon. Käytännössä pieniä yhteensopivuusongelmia esiintyy usein, ja muodon käyttö on vähenemään päin. Tästä syystä FSD on vähitellen luopumassa RTF-muodon käytöstä. ▪ Ei hyväksytty KDK:ssa säilytys- eikä siirtokelpoiseksi tiedostomuodoksi. ▪ Kansainvälisesti laajasti hyväksytty säilytys- tai siirtokelpoiseksi muodoksi (DANS, LoC, NAA, UKDA).

SEG-Y

Koko nimi:	SEG Y rev 1 Data Exchange format
Uusin versio:	1.0 (toukokuu 2002)
Avoimuus:	Avoin ja dokumentoitu, kansainvälisen työryhmän ylläpitämä
Yhteensopivuus:	Alaspäin yhteensopiva version rev 0 kanssa
Ohjelmistotuki:	Yleisesti tuettu seismologian ohjelmistoissa
Validointi:	Validaattoreita ei saatavilla
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/363
Aineistot:	FIRE-esimerkkiaineisto, seismologian aineistot
Huomioita:	▪ Seismologian alalla jo 1970-luvulta alkaen käytetty binäärisen datan tiedostomuoto.

- Tiedostomuoto sisältää vaihtoehtoisen ihmissilmin luettavan tekstimuotoisen header-osuuden, joka ei kuitenkaan FIRE-aineiston esimerkeissä ollut käytössä.
- Binäärinen header-osio, standardi määrittelee varsin paljon kenttiä joiden tiedot voidaan kirjata siihen (suurin osa vapaaehtoisia).
- Pitkäaikaissäilytyksen osalta määriteltävä, mitkä vapaaehtoiset kentät edellytetään täytettäväksi säilytettäviksi hyväksyttävissä tiedostoissa, sekä niiden sisältöön liittyvät yksityiskohdat. Validaattori olisi myös hyödyllinen, koska tiedostojen oikeellisuutta on silmämääräisesti mahdoton arvioida.
- Ei mainintoja kansainvälisillä suositeltavien tiedostomuotojen listoilla.

SIARD

Koko nimi:	Software Independent Archiving of Relational Databases
Uusin versio:	2.0
Avoimuus:	Avoin ja dokumentoitu, Sveitsin valtion ylläpitämä standardi
Yhteensopivuus:	Alaspäin yhteensopiva version 1.0 kanssa
Ohjelmistotuki:	Muodon kehittäneiden hankkeiden julkaisema avoimen lähdekoodin työkalu Database Preservation Toolkit [DBPTK]. Ei toistaiseksi tuettu muissa relaatiotietokantojen käsittelyyn tarkoitetuissa ohjelmissa.
Validointi:	Validaattori saatavilla (http://coptr.digipres.org/KOST-Val). Ei ilmeisesti vielä tue version 2.0 validointia.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000426.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/161 (versio 1.0)
Aineistot:	SMEAR-aineisto (muunnettu alkuperäisestä MySQL-muodosta)
Huomioita:	▪ Erityisesti relaatiotietokantojen pitkäaikaissäilytystä varten kehitetty XML-pohjainen tiedostomuoto. ▪ Kehitetty alun perin Sveitsin kansallisarkistossa, nykyisin mukana myös muita eurooppalaisia organisaatioita. ▪ Tukee kaikkia SQL:2008-standardin datatyyppejä ja olennaisimpia toimintoja. Eri relaatiotietokantojen valmistajakohtaiset ominaisuudet (erityisesti ohjelmointitoiminnot) eivät ole tuettuja. ▪ SIARD-tiedosto voi sisältää binäärisiä osia, mikäli tietokantaan on tallennettu binäärisiä objekteja ▪ Julkaistu avoimen lähdekoodin ohjelmisto tukee yleisimpiä käytössä olevia relaatiokantoja, siten että niiden sisältämät tiedot voidaan muuntaa SIARD-muotoon ja takaisin. Muunnos takaisin voi kohdistua myös muuhun kuin alkuperäiseen kantaan. ▪ Muunnostyökaluissa lyhyen testauksen perusteella vielä bugeja (erityisesti SIARDista takaisin muuntamisen osalta). ▪ Itse SIARD-muoto hyvin dokumentoitu ja näyttää olevan

vakiintumassa relaatiokantojen pitkäaikaissäilytyksen de facto –muodoksi.
- Hyväksytty säilytettäväksi muodoksi joissakin kansainvälisissä organisaatioissa (CINES, DANS).

SPSS Portable

Koko nimi:	Statistical Package for the Social Sciences (SPSS) Portable file format
Uusin versio:	24.0 (maaliskuu 2016, ohjelmiston versio) Tiedostomuoto ei ole muuttunut enää moniin vuosiin, tietoa viimeisestä muutosajankohdasta ei saatavilla.
Avoimuus:	Dokumentaatiota ei julkisesti saatavilla. FSD:llä on hallussaan SPSS Inc -yrityksen aikanaan toimittama dokumentaatio.
Yhteensopivuus:	Alaspäin ja ylöspäin yhteensopiva
Ohjelmistotuki:	Tuettu useimmissa kaupallisissa tilasto-ohjelmistoissa ainakin osittain, yhteensopivuusongelmia voi esiintyä
Validointi:	Validaattoreita ei saatavilla
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	FSD-esimerkkiaineisto, yhteiskuntatieteiden aineistot
Huomioita:	▪ Kaupallisen IBM SPSS -tilasto-ohjelman käyttämä muoto. ▪ Nimen "Portable" tarkoittaa siirrettävyyttä eri tietokonearkkitehtuurien välillä. ▪ Tuettu myös monissa muissa tilasto-ohjelmistoissa, joitakin yhteensopivuusongelmia esiintyy mm. merkistöjen osalta ▪ Pitkäaikaissäilytyksen kannalta haasteellinen dokumentaation ja tiedostomuotoa tukevien avoimen lähdekoodin ohjelmistojen puuttumisen takia. ▪ Käytännön testeissä todettu hyvin alas- ja ylöspäin yhteensopivaksi ja ollaan KDK:ssa alustavasti hyväksymässä säilytyskelpoiseksi tietyin reunaehdoin. ▪ Hyväksytty säilytettäväksi muodoksi joissakin kansainvälisissä organisaatioissa (DANS, UKDA).

TSV

Koko nimi:	Tab Separated Values (TSV)
Uusin versio:	Ei versiointia
Avoimuus:	Avoin, hyvin yksinkertainen muoto, ei varsinaista standardointia. Puolivirallinen muodon määrittelevä dokumentti saatavilla: https://www.iana.org/assignments/media-types/text/tab-separated-values

Yhteensopivuus:	TSV-tiedostot ovat periaatteessa keskenään sekä alas- että ylöspäin yhteensopivia, mutta koska vain kenttien erotin on määritelty, niissä voidaan käyttää mm. keskenään epäyhteensopivia merkistöjä
Ohjelmistotuki:	Tuettu monissa ohjelmissa, tuki helppo toteuttaa ohjelmoitaessa itse
Validointi:	Vain hyvin yksinkertainen validointi mahdollista määrityksen yksinkertaisuuden vuoksi: voidaan lähinnä tunnistaa onko kyseessä TSV-muotoinen tiedosto ja laskea onko joka rivillä sama määrä kenttiä
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	Aivokuvat- ja SMEAR-esimerkkiaineistot, todennäköisesti yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).
Huomioita:	▪ Helppo luoda ja käyttää, ihmissilmin ja koneellisesti luettavissa. ▪ Yksinkertaisuus kuitenkin haaste pitkäaikaissäilytyksen kannalta, koska tiedostomuoto jättää määrittelemättä seikkoja, joiden olisi hyödyllistä olla aineistojen kesken yhtenäisiä (mm. käytettävä merkistö). ▪ Tiedoston sisään ei voi standardoidusti tallentaa mitään metatietoja, joten pitkäaikaissäilytyksen osalta asia pitäisi ratkaista määrittelemällä tsv-tiedoston ohessa toimitettava, erillinen metatiedot sisältävä tiedosto. ▪ TSV-muotoiset aineistot ovat toisaalta helposti muunnettavissa KDK:ssa säilytyskelpoiseksi hyväksyttyyn CSV-muotoon, ja metatiedot toimitettavissa saman määrityksen mukaisesti ADDML-muodossa [KDK_Tiedostomuodot]. ▪ Hyväksytty säilytys- tai siirtokelpoiseksi muodoksi joissakin organisaatioissa (LoC, NAA, UKDA), muissa säilytettävissä tekstinä tai muunnettuna CSV-muotoon.

TXT (normaali)

Koko nimi:	Tekstitiedosto (Plain text, TXT)
Uusin versio:	Ei versiointia
Avoimuus:	Avoin, ei rakennetta joten ei myöskään dokumentaatiota tai standardointia.
Yhteensopivuus:	Tiedostot keskenään yhteensopivia, mikäli käytetty merkistö on sama.
Ohjelmistotuki:	Laajasti tuettu eri ohjelmissa
Validointi:	Ei mahdollista validoida. Käytetyn merkistön tarkistavia validaattoreita on saatavilla, mutta teknisistä syistä tarkistus ei ole kaikissa tapauksissa luotettava.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa

PRONOM: http://www.nationalarchives.gov.uk/pronom/x-fmt/111

Aineistot: FSD- ja Suomi24-esimerkkiaineistot, yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).

Huomioita:
- Ihmissilmin luettavaksi tarkoitettu tiedostomuoto, jota voidaan käyttää yksinkertaisten dokumenttien tallentamiseen ilman muotoiluja tai kuvia.
- Hyväksytty KDK:ssa säilytyskelpoiseksi tiedostomuodoksi, edellyttäen että käytetty merkistö on ISO 8859-15 tai UNICODE (UTF-8, UTF-16 tai UTF-32).
- Kansainvälisesti laajasti hyväksytty säilytyskelpoiseksi muodoksi (CINES, DANS, LAC, LoC, NAA, UKDA), tyypillisesti UNICODE- tai ASCII-merkistöllä.

TXT (rakenteinen)

Koko nimi: Tekstitiedosto (Plain text). Saattaa olla nimetty myös toisin, rakenteesta riippuen. Myös tiedostopääte saattaa vaihdella.

Uusin versio: Pääsääntöisesti ei versiointia.

Avoimuus: Avoin. Rakenne saattaa olla dokumentoitu tai dokumentoimaton.

Yhteensopivuus: Eri rakennetta käyttävät tiedostot eivät ole keskenään yhteensopivia.

Ohjelmistotuki: Laajasti tuettu eri ohjelmissa ihmissilmin tarkastelun ja manuaalisen muokkauksen osalta. Rakenteiden ohjelmistotuki suppeampi.

Validointi: Riippuu rakenteesta. Pääsääntöisesti validaattoreita ei ole saatavilla.

Eheys: Ei erityisiä mekanismeja eheyden säilyttämiseksi.

LoC-linkki: Ei kuvausta LoC-tiedostomuotokirjastossa

PRONOM: http://www.nationalarchives.gov.uk/pronom/x-fmt/111

Aineistot: ERNE-, FIRE- ja RITU-esimerkkiaineistot, yleisesti käytössä myös monissa muissa tutkimusaineistoissa, tyypillisesti tiedealakohtaisina, keskenään epäyhteensopivina muotoina.

Huomioita:
- Tekstitiedostoja käytetään tutkimusaineistoissa muotoilemattoman tekstin lisäksi usein myös erilaisten rakenteiden tallentamiseen.
- Rakenteet voivat olla esim. mittausparametreja, avain-arvo-pareja, taulukoita (joissa arvot erotettu toisistaan välilyönneillä) tai edellisten yhdistelmiä.
- Ihmissilmin luettavissa ja muokattavissa, yleensä myös helpohko käsitellä omissa ohjelmissa - valmiita ohjelmakirjastoja ei kuitenkaan tyypillisesti saatavilla rakenteiden keskinäisen erilaisuuden vuoksi.
- Jos joku tietty rakenne on tiedealalla laajemmin käytetty, sitä voidaan pitää omana tiedostomuotonaan (ks. esim. VRT).
- Rakenteelliset tekstitiedostot on hyväksytty KDK:ssa säilytyskelpoiseksi tiedostomuodoksi normaalina tekstinä, edellyttäen että käytetty merkistö on ISO 8859-15 tai UNICODE (UTF-8, UTF-16 tai UTF-32). Rakenne suositellaan kuvattavaksi

ADDML-metatietoskeemalla.

- Joissain tapauksissa rakenteisten tekstitiedostojen muuntaminen paremmin koneluettavaan muotoon saattaa olla järkevää. Esimerkiksi taulukot voisi muuntaa CSV-muotoon ja avain-arvo-parit JSON-muotoon.
- Kansainvälisesti laajasti hyväksytty säilytyskelpoiseksi muodoksi normaalina tekstinä (CINES, DANS, LAC, LoC, NAA, UKDA), tyypillisesti UNICODE- tai ASCII-merkistöllä. Ei erikseen ohjeita rakenteiseen tekstiin liittyen.

VRT

Koko nimi:	Verticalized Text (VRT). (Corpus Workbench-ohjelman käyttämä muoto, ei GDAL - Geospatial Data Abstraction Library Virtual Format, joka käyttää samaa lyhennettä VRT.)
Uusin versio:	Ei tiedossa / ei versiointia
Avoimuus:	Avoin. Dokumentaatio puutteellinen.
Yhteensopivuus:	Ainakin pääosin alas- ja ylöspäin yhteensopiva.
Ohjelmistotuki:	Tuettu IMS Open Corpus Workbench -ohjelmassa (http://cwb.sourceforge.net/) ja vaihtelevasti kielitieteilijöiden kehittämissä omissa ohjelmissa
Validointi:	Validaattoria ei saatavilla.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	Ei kuvausta LoC-tiedostomuotokirjastossa
PRONOM:	Ei kuvausta PRONOM-tiedostomuotokirjastossa
Aineistot:	Suomi24-esimerkkiaineisto
Huomioita:	▪ Tekstitiedosto, jonka rakenne muistuttaa XML:ää, kuitenkin joiltain osin siitä poiketen. Rakenne sisältää myös taulukoita, joiden kentät on erotettu toisistaan välilyönneillä. ▪ Rakenne itsessään on melko selkeä ja helposti ymmärrettävissä, mutta siinä käytetään vakioituja lyhenteitä, joita ei ole dokumentoitu. Kielitieteen asiantuntija osaa haastatellun asiantuntijan mukaan päätellä, mitä tagit ja lyhenteet tarkoittavat. ▪ Voitaisiin periaatteessa säilyttää KDK:ssa säilytyskelpoiseksi hyväksyttynä tekstitiedostona, mutta ymmärrettävyyden säilymisen varmistamiseksi sekä rakenne että käytetyt lyhenteet tulisi dokumentoida. ▪ Ei mainintoja tarkasteltujen ulkomaisten organisaatioiden suositeltavien tiedostomuotojen listoilla.

WMV

Koko nimi:	Windows Media Video (WMV)
Uusin versio:	WMV 9
Avoimuus:	Tiedostomuodon versio 9 on avoin, dokumentoitu ja standardoitu (SMPTE 421M). Tiedostoissa on kuitenkin mahdollisuus käyttää salausta ja Digital Rights Management (DRM) -lisäosia, jotka eivät ole osa standardia eivätkä myöskään avoimia.
Yhteensopivuus:	Alaspäin ja ylöspäin yhteensopiva
Ohjelmistotuki:	Laajasti tuettu eri ohjelmissa (standardoitu, salaamaton versio)
Validointi:	Varsinaisia validaattoreita ei saatavilla. Tiedoston eheys on kuitenkin mahdollista osittain tarkistaa lukemalla se jollain muotoa tukevalla ohjelmalla (esim. ffmpeg, https://www.ffmpeg.org/) ja tarkistamalla, tuleeko luettaessa virheitä.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000091.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/133
Aineistot:	Aivokuvat- ja FIRE-esimerkkiaineistot, todennäköisesti yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).
Huomioita:	▪ Häviöllistä pakkausta käyttävä videotiedostomuoto. ▪ Versio 9 hyväksytty KDK:ssa siirtokelpoiseksi tiedostomuodoksi [KDK_Tiedostomuodot]. ▪ Vanhempia versioita tai standardiin kuulumattomia DRM-lisäosia ei tule käyttää säilytettäviksi toimitettavissa tiedostoissa. ▪ Kansainvälisesti hyväksytty joissakin organisaatioissa (LAC, NAA).

XLSX

Koko nimi:	Office Open XML Spreadsheet (XLSX)
Uusin versio:	ISO/IEC DIS 29500 (2012)
Avoimuus:	Dokumentoitu ja standardoitu.
Yhteensopivuus:	Alaspäin ja ylöspäin yhteensopiva
Ohjelmistotuki:	Tuettu useissa eri ohjelmissa. Kaikkien ominaisuuksien täysin toimiva tuki vain Microsoft Excelissä.
Validointi:	Validaattoreita ei saatavilla.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000398.shtml

PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/214
Aineistot:	Kiteet-esimerkkiaineisto, todennäköisesti yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).
Huomioita:	▪ Microsoft Excel -taulukkolaskentaohjelman käyttämä tiedostomuoto, joka on vähintään osittain tuettu myös monissa muissa ohjelmissa. ▪ Hyväksytty KDK:ssa siirtokelpoiseksi tiedostomuodoksi. ▪ Kansainvälisesti laajasti hyväksytty siirtokelpoiseksi muodoksi (DANS, LAC, LoC, NAA, UKDA).

XML

Koko nimi:	Extensible Markup Language (XML)
Uusin versio:	XML 1.0 Fifth Edition (marraskuu 2008) - yleisimmin käytetty muoto XML 1.1 Second Edition (elokuu 2008) - erityistarkoituksiin, joissa tarvitaan version 1.1 uusia ominaisuuksia
Avoimuus:	Avoin, dokumentoitu ja standardoitu
Yhteensopivuus:	Alas- ja ylöspäin yhteensopiva
Ohjelmistotuki:	Laajasti tuettu eri ohjelmissa.
Validointi:	Useita validaattoreita saatavilla.
Eheys:	Ei erityisiä mekanismeja eheyden säilyttämiseksi.
LoC-linkki:	http://www.digitalpreservation.gov/formats/fdd/fdd000075.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/101
Aineistot:	FSD- ja RITU-esimerkkiaineistot, todennäköisesti yleisesti käytetty myös monissa muissa tutkimusaineistoissa (tiedealariippumaton muoto).
Huomioita:	▪ Merkintäkieli, jota voidaan käyttää sekä dokumentaation, metatietojen että datan tallentamiseen. ▪ Käytetty rakenne voidaan määritellä formaalisti XML-skeemojen avulla. ▪ Luettavissa sekä ihmissilmin että koneellisesti. ▪ Versio 1.0 hyväksytty KDK:ssa säilytyskelpoiseksi tiedostomuodoksi [KDK_Tiedostomuodot]. ▪ Kansainvälisesti laajasti hyväksytty säilytyskelpoiseksi muodoksi (CINES, DANS, LoC, NAA, UKDA).

LIITE D. KDK-PASIN SÄILYTYSKELPOISTEN TIEDOSTOMUOTOJEN VALINTAKRITEERIEN SOVELTUVUUS TUTKIMUSAINEISTOILLE

Lähtökohdat

Lähtökohtana säilytyskelpoisten tiedostomuotojen valinnassa ovat KDK-PAS:n säilytyskelpoisiksi hyväksyttyjen tiedostomuotojen valintakriteerit, jotka on esitetty säilytys- ja siirtokelpoiset tiedostomuodot esittelevän dokumentin liitteessä B [KDK_Tiedostomuodot]. Ne ovat avoimuus, käyttö PAS-standardina, vakaus / yhteensopivuus, riippuvuudet / yhteentoimivuus ja standardisuus. Tässä osiossa arvioidaan, miltä osin tutkimusaineistojen erityispiirteet edellyttävät täydennyksiä tai muutoksia kriteereihin.

Avoimuus

Ideaalitilanteessa tiedostomuodon määrityksen on luonut ja niitä jakelee standardointijärjestö tai muu avoimen jäsenyyden kansainvälinen järjestö. Tutkimusaineistojen kohdalla on kuitenkin varsin yleistä, että määritys on sinänsä avoin, mutta sen on luonut joko yksittäinen yliopisto tai alan tutkijoiden epävirallinen yhteistyöelin, joka ei ole varsinainen järjestö. Määritykset ovat lähes aina saatavilla maksutta, mutta mahdollisesti vain yhdestä paikasta. Toisaalta niiden kopiointi on yleensä sallittua.

Standardoituja tiedostomuotoja on hyvä suosia, mutta olennaisin hyväksymiskriteeri tulee olla se, että määritys on ylipäätään avoimesti saatavilla. Se mahdollistaa aineiston hyödyntämisen myös alkuperäisestä täysin poikkeavissa käyttötarkoituksissa, esimerkiksi datan analysoinnin uusin menetelmin tutkijan itse kehittämällä ohjelmalla. Määrityksen sijaintipaikkojen lukumäärä ei ole kovin olennainen. Koska määrityksen pitkäaikaista saatavuutta ei erityisesti epävirallisten yhteistyöelimien luomien tiedostomuotojen osalta voida taata, tulee siitä tallentaa kopio PAS-ratkaisussa säilytettävän aineistokokonaisuuden yhteyteen.

Tutkimusaineistoissa on myös varsin runsaasti tutkijoiden tai tutkimusryhmien itse luomia tai käytettyyn mittalaitteeseen liittyviä tiedostomuotoja, joille ei ole julkaistu määritystä. Niiden osalta tulee edellyttää, että tiedostomuodon kuvaava dokumentti laaditaan ennen kyseisen muodon hyväksymistä säilytyskelpoiseksi.

Käyttö PAS-standardina

KDK-PAS:n arvio perustuu siihen, kuinka moni kulttuurialan organisaatioista käyttää tai aikoo käyttää tiedostomuotoa säilytyskelpoisena tiedostomuotona. Tämä ei sovellu tutkimusaineistojen arviointiin, koska yliopistoilla ei ole kirjastojen, museoiden ja arkistojen tapaan aineistojen säilytysvelvollisuutta eivätkä ne ole arvioineet tiedostomuotojen säilytyskelpoisuutta.

Samankaltaisena kriteerinä voitaisiin periaatteessa käyttää sitä, kuinka monessa data-arkistossa tiedostomuoto on kansainvälisesti hyväksytty säilytyskelpoiseksi. Käytännössä tähän on hankala tukeutua, koska vain harvat organisaatiot ovat julkaisseet listoja hyväksymistään tiedostomuodoista, eivätkä listat ole kattavia. Lisäksi säilytyksen tasot vaihtelevat, esimerkiksi sen osalta kuinka pitkäksi aikaa organisaatio sitoutuu aineistojen säilytykseen, sekä kuinka paljon ymmärrettävyyden säilytykseen on kiinnitetty huomiota.

Vakaus / yhteensopivuus

Vakauteen ja yhteensopivuuteen liittyvät KDK-PAS:n arviointikriteerit soveltuvat periaatteessa myös tutkimusaineistojen tiedostomuodoille. Käytännössä esimerkiksi tiedostomuotojen alas- ja ylöspäin yhteensopivuudesta on vaikea saada luotettavaa tietoa. Versiopäivitysten määrä ja uusimman käytössä olevan version ikä on yleensä varsin helppo selvittää.

Korruptoitumisen sietokyky riippuu tiedostomuodon lisäksi myös analyysimenetelmästä. Joissakin tapauksissa pienikin muutos voi tehdä tiedostosta täysin käyttökelvottoman, kun taas toinen menetelmä sietää poikkeamia paremmin. Joillakin tutkimusaloilla on otettu käyttöön vanhoja tiedostomuotoja huonommin korruptoitumista sietäviä uusia muotoja, lähinnä siksi että ne tallentavat dataa tiiviimmin ja vaativat siten vähemmän tallennuskapasiteettia. Erityisen suurten aineistojen tapauksessa kustannussäästö voi olla merkittävä.

Korruptoitumisen sietokyky osana tiedostomuodon rakennetta ei ole erityisen tärkeä ominaisuus, koska PAS-ratkaisu huolehtii joka tapauksessa tiedostojen eheydestä sen jälkeen, kun ne on saatu siirrettyä säilytykseen.

Riippumattomuus, yhteentoimivuus

Riippumattomuuden ja yhteentoimivuuden arviointi perustuu siihen, missä määrin tiedostomuoto on sidottu tiettyyn laitteistoon tai ohjelmistoon. KDK-PAS:n arviot eivät mainitse tiettyjä lukumääriä, vaan käyttävät termejä "korkea", "keskitason" ja "alhainen" riippumattomuus tai yhteentoimivuus.

Kriteeri soveltuu varsin hyvin myös tutkimusaineistojen tiedostomuotojen arviointiin. Tiedostomuodon tuki vähintään kahdessa eri ohjelmassa on merkittävä etu riippumattomuuden ja yhteentoimivuuden kannalta. Se viittaa myös siihen, että tiedostomuodon rakenteen kuvaava määritys on riittävä tiedostojen siirtämiseen eri ohjelmien välillä. Toisaalta tuen tasoa eri ohjelmissa on vaikea arvioida ilman syvällistä perehtymistä. Vaikka tiedostomuoto olisi listattu tuetuksi, saattaa olla että ohjelma tukee vain osaa sen ominaisuuksista.

Arvioinnissa on perusteltua huomioida, ovatko tiedostomuotoa tukevat ohjelmat avoimen lähdekoodin ohjelmia. Avoimena lähdekoodina ja uudelleenkäytön sallivalla lisenssillä toteutettu tuki on suljettua arvokkaampi, koska sitä voi hyödyntää pohjana myös uuden analyysiohjelman itse kirjoittava tutkija.

Standardisuus

Standardisuuden arvio perustuu KDK-PAS:ssa siihen, millaisella prosessilla tiedostomuotoa säännellään. Tämä soveltuu tutkimusaineistojen tiedostomuotojen arviointiin vain hyvin rajoitetusti. Valtaosan tiedostomuodoista osalta prosessia ei ole määritelty, vaan muotoa käytetään niin kauan kuin se palvelee tutkimusyhteisöä hyvin. Kun uudet tutkimusmenetelmät edellyttävät muutoksia, tutkijat tekevät usein laajennuksia itse. Yhteiskäyttöön soveltuva päivitetty versio tiedostomuodosta saatetaan laatia esimerkiksi laajennuksen laatineen tutkimusryhmän tekemän ehdotuksen pohjalta tai merkittävän alan konferenssin yhteydessä kerätyn palautteen perusteella.

Arvioon voitaisiin sen sijaan sisällyttää se, kontrolloiko tiedostomuotoa kaupallinen toimija, tutkimusorganisaatio (esim. yliopisto) vai tutkimusyhteisö. Kahdessa jälkimmäisessä tapauksessa on todennäköisempää, että tiedostomuodon kehitys palvelee kansainvälisen tutkimusyhteisön etua. Koska yhteistyötä eri tutkimusryhmien välillä tehdään yhä enemmän, tutkimusaineistoissa on alasta riippumatta havaittavissa selvä suuntaus yhteisesti hyväksyttyjen tiedostomuotojen käyttöön.