AVOIN TIEDE
JA TUTKIMUS

TUTKIMUKSEN PAS-PALVELUN PILOTIT 2015: LOPPURAPORTTI

Julkaisu Tutkimuksen PAS-palvelun pilotit 2015: Loppuraportti	
Julkaisija Avoin tiede ja tutkimus –hanke	Julkaisuajankohta 8.4.2016
Tekijä Tutkimus-PAS -työryhmä	
Lisenssi	
Julkaisun jakelun PDF-tiedosto ladattavissa sivuilla avointiede.fi/keskeiset-julkaisut	
Yhteystiedot http://avointiede.fi avointiede@postit.csc.fi	

SISÄLLYS

1 JOHDANTO

Tämä dokumentti on Avoin tiede ja tutkimus -hankkeen (ATT-hanke) tutkimuksen pitkäaikaissaatavuuspalvelun (Tutkimuksen PAS-palvelun) vuoden 2015 pilottien loppuraportti, joka kokoaa yhteen vuoden aikana toteutettujen pilottien tulokset. Pilottien keskeinen tarkoitus oli tuottaa tietoa Tutkimuksen PAS-palvelun vaatimusmäärittelyn laatimisen tueksi ja parantaa pilottiin osallistuvien tutkimusorganisaatioiden teknisiä valmiuksia PAS-palvelun käyttöön. Piloteissa syvennettiin aiemmin vuoden 2014 aikana toteutetuista piloteista saatuja kokemuksia, ja hyödynnettiin Kansallisen digitaalisen kirjaston pitkäaikaissäilytyspalvelun (KDK:n PAS-palvelun) teknisiä ratkaisuja ja määrityksiä.

Pitkäaikaissaatavuuden edellytyksenä on pitkäaikaissäilytys, ja näissä piloteissa keskityttiin ensisijaisesti tutkimusaineiston ymmärrettävyyden säilytyksen kannalta olennaisten kysymysten ratkaisemiseen. Bittitason säilyttämisessä voidaan pitkälti nojautua KDK:n PAS-palvelun kanssa yhteiseen PAS-ratkaisuun, joten sen pilotoimiselle ei ollut tarvetta.

Kussakin pilotissa tutkimusorganisaatio valitsi pilotoitavaksi soveltuvan tutkimusaineiston. Tavoitteena oli selvittää millaisia vaatimuksia valitun aineiston säilyttäminen asettaisi Tutkimuksen PAS-palvelulle. Pilottien käytännön toteutus rakentui kolmen tarkastuspisteen varaan siten, että kuhunkin tarkastuspisteeseen asetettiin tavoitteet, jotka aikataulutettiin yhdessä tutkimusorganisaation kanssa. Tarkastuspisteet olivat 1) metatietojen ja tiedostomuotojen tunnistaminen, 2) aineiston paketointi ja siirtäminen sekä 3) yhteenveto ja raportointi.

Ensimmäisessä tarkastuspisteessä keskityttiin ymmärrettävyyden säilyttämisen kannalta olennaisten metatietojen ja aineiston tiedostomuotojen arviointiin. Olennaisia metatietoa ovat esimerkiksi tutkimusaineiston sisältöä kuvaava kuvaileva metatieto, aineiston hallintaan liittyviä tietoja kuvaava hallinnollinen metatieto ja tutkimusaineiston rakennetta kuvaava rakenteellinen metatieto. Metatietoja kartoitettiin metatietokyselyllä (liite A), jossa lähtökohtana oli selvittää kaikki aineistoon liittyvät olennaiset metatiedot ja tämän jälkeen arvioida niiden merkitys ymmärrettävyyden säilyttämisen kannalta. Tämän lisäksi käytössä olevien tiedostomuotojen osalta arvioitiin, mitkä niistä ovat mahdollisimman pysyviä ja säilyttävät parhaiten aineiston tekniset ominaisuudet, mikä on niiden dokumentaation taso ja toisaalta mitkä olivat mahdollisia vaihtoehtoisia tiedostomuotoja.

Aineistojen paketoinnissa lähtökohtana käytettiin KDK:n PAS-palvelun paketointimääritystä[1] (KDK:n hallinnolliset ja rakenteelliset metatiedot ja aineiston paketointi, v. 1.4). Paketoitu aineisto siirrettiin KDK:n PAS-palvelun testiympäristöön, jossa sille ajettiin KDK:n määritysten mukaiset tarkastukset. Paketoinnin tarkoituksena oli tutkimusaineistojen vaatimusten tunnistaminen ja erityisesti vaatimusten eroavaisuuksien tunnistaminen suhteessa kulttuuriaineistoihin.

Pilotteja toteutettiin yhteensä kolme kappaletta[2] seuraavien organisaation kanssa:

- Aalto-yliopiston Brain & Mind Laboratory (BML)
- Jyväskylän yliopiston Kiihdytinlaboratorio
- Turun yliopiston Avaruustutkimuslaboratorio (ATL)

[1] http://www.kdk.fi/fi/pitkaaikaissailytys/maeaerittely-ja-dokumentit/pitkaeaikaissaeilytys/145-kdkn-hallinnolliset-ja-rakenteelliset-metatiedot-ja-aineiston-paketointi

[2] Alkuperäisenä suunnitelmana oli toteuttaa pilotteja neljä kappaletta, mutta tästä suunnitelmasta poiketen pilotti Suomen molekyylilääketieteen instituutin (Institute for Molecular Medicine Finland, FIMM) kanssa ei valmistunut aikataulussa. Tässä pilotissa aineistona käytettiin 1000 Genomies -projektin *"An integrated map of genetic variation from 1,092 human genomes"*. Ihmisen genomitiedon pitkäaikaissaatavuuteen liittyviä kysymyksien on tarkoitus selvittää näistä piloteista erillisenä vuoden 2016 aikana.

Keskimäärin pilotin ajallinen kesto oli noin 4–5 kuukautta, mutta tämä piti sisällään myös teknistä valmistelua. Tutkimusorganisaation osalta työmäärä oli keskimäärin noin 3–4 henkilötyökuukautta.

Pilottien organisoinnista vastasi Tutkimus-PAS-työryhmä, ja toteutuksesta Tutkimus-PAS-projektiryhmä, jossa vastuuorganisaatioina ovat Kansalliskirjasto ja CSC – Tieteen tietotekniikan keskus Oy. Tämä raportti on kirjoitettu yhteistyössä Tutkimus-PAS-projektiryhmän ja pilotteihin osallistuneiden tutkimusorganisaatioiden kanssa.

2 PILOTTI: AALTO-YLIOPISTON BRAIN & MIND LABORATORY

2.1 Pilotin tiedot

Aineiston nimi	Elokuvan käyttö aivotutkimuksessa
Aineiston koko	8 gigatavua (sisältäen mittausaineiston, mittalaitteen konfiguraation, potilastiedot, ärsykkeenä käytetyn elokuvan ja eettisen toimikunnan luvan)

Aalto-yliopiston Brain & Mind Laboratory (BML) osallistui Tutkimuksen PAS-palvelun pilotointiin syksyn 2015 aikana. BML on kansainvälinen Aalto-yliopiston neurotieteen ja lääketieteellisen tekniikan laitoksen tutkimusryhmä, jossa tehdään tutkimusta aivo- ja hermomekanismien vaikutuksesta ihmisen mieleen. Tutkimuksessa karakterisoidaan käyttäytymistä, arvioidaan subjektiivisia kokemuksia, kerätään taustatietoja potilaista ja mitataan aivojen toimintaa joko toiminnallisella magneettikuvauksella (fMRI), magnetoenkefalografialla (MEG) tai aivosähkökäyrällä (EEG).

Pilotoitavaksi aineistoksi valittiin “Elokuvan käyttö aivotutkimuksessa”, jossa koehenkilöiden aivoja kuvattiin toiminnallisella magneettikuvauksella, samalla kun heille näytettiin valittua elokuvaa. Toiminnallisessa magneettikuvauksessa tutkitaan ihmisen aivoja luonnollisissa olosuhteissa eli esimerkiksi äänen tai liikkuvan kuvan ärsykkeiden vaikutuksessa. Tämä on erityisen hyödyllistä siksi, että tällöin mittauksia pystytään tarkasti analysoimaan ja vertaamaan ärsykkeeseen. Pilotoitavaksi valittu aineisto on ensimmäinen luonnollinen toiminnallinen magneettikuvausaineisto, joka on tehty BML:n laboratoriossa. Ajallisesti pilotti kesti kolme kuukautta (loka–joulukuu) aikana, ja siihen osallistuivat prof. Mikko Sams (projektin johto), TkT Enrico Glerean (päätoimisesti) ja konsultointiapuna ryhmän muut aineiston keräyksessä mukana olleet tutkijat.

Pilotissa haluttiin varmistaa tärkeiden tutkimusaineistojen säilyttäminen tulevaisuudessa ja aineistoon liittyvän kuvailun tallentaminen vakioidulla tavalla. BML:llä oli aiempaa kokemusta säilyttämiseen ja metatietoihin liittyvistä kysymyksistä mutta ei kuitenkaan kokemusta pitkäaikaissäilyttämisestä tai siihen liittyvistä kysymyksistä.

2.2 Pilotin toteuttaminen

2.2.1 Metatiedot

Aineiston metatiedot olivat hyvin selvillä, koska keskeiset metatiedot oli tallennettu aineiston yhteyteen ja koska kokeen suorittajat olivat vielä tavoitettavissa. Kaikkiin metatietoihin liittyviin kysymyksiin löydettiin lopulta vastaukset, vaikka osa edellyttikin yhteyden ottamista alkuperäisten mittauksen suorittajiin. Aineisto oli arkistoitu BML:n käyttämän vakiomuotoiseen rakenteeseen, joten oli ennalta tiedossa mistä mikäkin tieto löytyy.

Metatietokyselyn keskeiset tulokset koskivat aineiston käyttöoikeuksia ja saatavuuden varmistamista. Aineisto on tuotettu lääketieteellisessä kokeessa ja sen tuottamiseen on pyydetty lupa eettiseltä

toimikunnalta (Helsingin ja Uudenmaan sairaanhoitopiirin Lasten ja nuorten sairauksien ja psykiatrian eettinen toimikunta). Aineisto on mittauksen jälkeen anonymisoitu. Eettisen toimikunnan lupa kattaa aineiston tuottamisen, mutta juridisesti kysymys aivokuvantamisaineistojen julkaisemisesta on epäselvä. Tällä hetkellä BML päättää aineiston jakamisesta eteenpäin tapauskohtaisesti: "*The dataset cannot be shared publicly. Collaborators can obtain access to the data by contacting the authors or owner of the dataset*."

Suomessa ei ole lääketieteellisten aineistojen asemasta yhtenäistä käytäntöä. Esimerkiksi Alzheimer-tutkimuksessa on muodostunut erillinen komitea (Alzheimer data initiative), joka myöntää pyynnöstä käyttöoikeuksia tutkimusaineistoihin. Pitkällä aikavälillä tilanteen odotetaan selkiytyvän lainsäädännön muutoksien kautta.

Aineistoon kuuluu kiinteäksi osaksi myös hieman muokattu versio valitusta elokuvasta. Elokuvan käyttöön osana tutkimusta on pyydetty tuottajan lupa ja elokuvaa tarvitaan aineistoa analysoitaessa. Elokuvaa ei kuitenkaan ole sallittua jakaa täysin vapaasti.

Pilotissa selvitettiin aivokuvantamisaineistojen kuvailuun käytössä olevia kuvailukäytäntöjä ja metatietostandardeja. Erityisesti aineiston löydettävyyden kannalta keskeisiksi metatiedoiksi tunnistettiin mittauksen parametrit, kuten mittausresoluutiot ja -tilavuudet sekä magneettikenttien voimakkuudet. Mittauksen parametrit tulevat suoraan mittalaitteelta ja ovat tällöin yleensä mittalaitekohtaisessa formaatissa. Parametrit ovat keskeisiä hakukriteerejä aineistoja etsittäessä.

2.2.2 Aineiston tekniset tiedot

BML:llä oli aineistojen arkistoimiseksi käytössä itse kehitetty vakiomuotoinen rakenne, jossa jokaiselle kokeeseen kuuluvalle tiedostolle oli määrätty paikka. Aivokuvantamisaineistojen paketoimiseksi vakiomuotoisiin rakenteisiin on tyypillistä. Alalla on tällä hetkellä menossa kaksi rakenteen määrittelemiseen tähtäävää hanketta: OpenfMRI- ja BIDS-rakenne. Pilotissa tehtiin työkalu, joka muuntaa BML:n omassa rakenteessa olevia aineistoja BIDS-rakenteen mukaiseen muotoon.

On nähtävissä, että BIDS-rakenteella tulee olemaan alalla merkittävä rooli. BIDS-rakenteen määritys on vielä luonnosasteella, mutta koska se perustuu aiempaan OpenfMRI:ssä tehtyyn työhön, uskottiin sen olevan paras tapa jäsentää aivokuvantamisaineistoja. BIDS-rakennetta varten on olemassa avoimena lähdekoodina kehitetty validaattorityökalu. Pilotissa BIDS-rakenteeseen muunnetut aineistot validoitiin tällä työkalulla.

BIDS-rakenteeseen kuuluu alkuperäiset magneettikuvausaineistot (fMRI) NIfTI-formaatissa, metatiedot JSON-muodossa, potilastiedot TSV-tekstimuodossa (tabular separated value) ja eettisen toimikunnan lupa PDF-muodossa. Säilytyksen kannalta nämä ovat kaikista keskeisimmät tiedostomuodot, sillä ne sisältävät aineiston raakamuodossa, tiedot koehenkilöistä ja itse mittaukseen liittyvät parametrit. Lisäksi BIDS-rakenteeseen on mahdollista sisällyttää esimerkiksi versionhallintatietoja tekstimuodossa.

Kokeessa käytetty videotiedosto on MPEG-4 muodossa, joka on säilytyskelpoinen tiedostomuoto KDK:n PAS-palvelussa.

2.2.3 Paketointi

Pilotoitavaksi valittu aineisto paketoitiin KDK:n PAS-palvelun määritysten mukaiseen siirtopakettiin. Siirtopaketti sisälsi aineiston BIDS-rakenteessa ja pilotissa kehitettiin tapa kuvata BIDS-rakenne ja sen metatiedot siirtopaketissa. Prosessi joka muuntaa BML:n aineistoja BIDS-rakenteeseen ja siitä edelleen siirtopaketeiksi, automatisoitiin. Paketoinnissa hyödynnetty työkalu julkaistiin avoimena lähdekoodina.

Paketoinnin suurimmat haasteet liittyivät BIDS-rakenteen metatietojen esittämiseen siirtopaketissa. Siirtopaketin rakenteesta johtuen (BIDS-rakenne siirtopaketissa) metatiedot joudutaan siirtopaketissa

esittämään uudestaan. Tämä on välttämätöntä, koska siirtopaketissa tulee esittää kaikki säilyttämisen kannalta olennaiset metatiedot.

Aineistoon liittyvät kuvailevat metatiedot kuvattiin siirtopaketissa Dublin Core -metatietoskeeman avulla. Aineistoon liittyvien teknisten metatietojen, lähinnä mittalaitteen parametrien kuvaaminen, oli kuitenkin ongelmallista, sillä siihen ei ole alalla yleisesti käytössä olevaa metatietoskeemaa.

Kokonaisuudessaan paketointi oli lähinnä tekninen prosessi, jossa ei ollut suuria ongelmia.

2.3 Yhteenveto

Pilotin suurimmat haasteet olivat lääketieteellisen aineistoon liittyvät käyttöoikeuskysymykset. Tekninen haaste oli nykyisen rakenteen muuntaminen BIDS-rakenteeseen ja tämän prosessin automatisoiminen. Aikaa vievää, mutta välttämätöntä työtä olivat alalla käytössä olevien kuvailukäytäntöihin tutustuminen ja niiden arviointi.

Pilotissa onnistuttiin kartoittamaan aineiston säilyttämisen ja saatavuuden kannalta olennaiset metatiedot. Kaikkia mittaukseen liittyviä parametreja ei onnistuttu kuvaamaan paketoinnissa; osa näistä voi tulevaisuudessa olla merkittäviä aineiston löydettävyyden kannalta. Tämä johtui siitä, että mittalaitekohtaisten parametrien kuvaamiseen ei löydetty soveltuvaa metatietostandardia. Tieto kuitenkin on säilytetty BIDS-rakenteen sisään, joten metatietoja voidaan tämän pohjalta täydentää tulevaisuudessa. Jotta aineisto voitaisiin vastaanottaa Tutkimuksen PAS-palvelussa, tarvittaisiin tuki BIDS-rakenteessa käytetyille tiedostomuodoille (NIfTI, JSON, TSV).

Merkittävin hyöty tutkimusryhmälle tästä pilotista oli ohjelmiston kehittäminen, joka muuntaa käytössä olevan hakemistorakenteen BIDS-rakenteeksi. Alalla ollaan selvästi siirtymässä BIDS-rakenteen käyttöön sekä aineiston käytössä että sen säilyttämisessä. Aineistoa on helpompi jakaa edelleen, ja toisaalta myös työkalut tukevat sen käyttöä.

Kokonaisuudessaan pilotti onnistu hyvin, se oli organisoitu hyvin ja Tutkimus-PAS-projektiryhmä tarjosi teknistä tukea tarvittaessa. BML on kiinnostunut osallistumaan palvelun käyttöön sen tuotantovaiheen alkaessa. Tarkoituksena on myös jakaa tietoa ja pilotin kokemuksia oman yliopiston sisällä ja muille neurotieteiden tutkijoille.

3 PILOTTI: JYVÄSKYLÄN YLIOPISTON KIIHDYTINLABORATORIO

3.1 Pilotin tiedot

Aineiston nimi	250-Nobeliumin hajoamisspektroskopia
Aineiston koko	200 gigatavua (sisältäen mittausaineiston, mittalaitteen lähdekoodit, julkaisut)

Pilottiin osallistui Jyväskylän yliopiston fysiikan laitoksen kiihdytinlaboratorio. Laboratorio on kansallisesti ja kansainvälisesti merkittävä tutkimusinfrastruktuuri, jossa tehdään pääosin ydinfysiikan perustutkimusta. Pitkäaikaissäilytyksen tarve perustuu ensisijaisesti siihen, että kokeissa käytetyiden laitteiden ja materiaalien saatavuutta tulevaisuudessa ei voida välttämättä taata. Kokeiden valmistelu ja suorittaminen on usein myös varsin arvokasta taloudellisesti ja datamäärät erittäin suuria, jopa kymmeniä teratavuja viikossa.

Pilottiin valittiin aineistoksi koe *"250-Nobeliumin hajoamisspektroskopia"*. Koe tutkii nimensä mukaisesti superraskaiden alkuaineiden radioaktiivisia hajoamistapahtumia käyttäen hyväksi kiihdytinlaboratorion

RITU-rekyyliseparaattoria ja GREAT-spektrometriä. Aineiston valintaperusteena käytettiin lähinnä mittausten läheistä ajankohtaa (maaliskuu 2015) ja verrattain pientä kokoa (200GB).

Pilotti toteutettiin touko–marraskuussa 2015. Pilotin toteutukseen osallistui aktiivisesti yksi yliopistotutkija jonka työtehtäviin kuului jo aiemmin aineistojen säilyttämisestä vastaaminen (FT Panu Rahkila). Laboratorion henkilökuntaa ja sidosryhmien edustajia konsultoitiin tarvittaessa koko pilotin ajan. Touko–elokuun aikana pilotointiin käytettiin lähinnä metatietokyselyyn vastaamiseen ja kuvailukäytäntöjen kartoittamiseen noin yksi henkilökuukausi. Syys–marraskuun ajalla paketointiohjelmiston kehittämisen ja teknisen paketoinnin aikana yksi henkilö työskenteli pilotissa päätoimisesti kahden kuukauden ajan opetus- ja kulttuuriministeriön rahoituksen tuella.

Kiihdytinlaboratoriolla oli myös aiempaa kokemusta pitkäaikaissäilyttämisestä, sillä tuotetut aineistot on pääosin yritetty pitkäaikaissäilyttää arkistoimalla aineistot magneettinauhoille. Tätä taustaa vasten tekniset valmiudet pilottiin olivat hyvät: kaikki tarvittavat aineistot löytyivät digitaalisessa muodossa keskitetyistä tallennusjärjestelmistä. Toisaalta etukäteen tiedettiin metatietoihin, käyttöoikeuksiin ja varsinkin aineistojen omistajuuteen liittyvät kysymykset ongelmallisiksi.

3.2 Pilotin toteuttaminen

3.2.1 Metatiedot

Metatietojen selvittämisen lähtökohtana oli aluksi selvittää kaikki aineistosta tiedossa olevat metatiedot ja tämän jälkeen arvioida niiden merkitys säilyttämisen kannalta. Metatietokyselyn avulla nämä voitiin kartoittaa kattavasti. Erityisen keskeisiksi metatiedoiksi tunnistettiin muuttujatiedot, jotka ovat erityisen keskeisiä aineiston löydettävyyden kannalta.

Aineiston kuvailun osalta käytiin läpi alan kansainväliset metatietostandardit, sanastot ja luokittelut. Alan keskeisenä toimijana CERN:n tutkimuskeskuksella on merkittävä rooli alan käytäntöjen muokkaamisessa, ja heidän kanssaan tehtiinkin asiassa yhteistyötä. Yhteistyöstä selvisi, että heidän käytössään oleva arkistointi perustuu kirjastoalan MARC-standardiin, jonka he ovat todenneet huonosti toimivaksi heidän käyttötarkoitukseensa. He ovat kehittämässä uutta kuvailukäytäntöä ydinfysiikan aineistojen kuvailuun, mutta työstä ei valmistunut käyttökelpoista versiota tämän pilotin tarpeisiin.

Aineiston kuvailussa käytettiin hyväksi American Institute of Physicsin laatimaa PACS (Physics and Astronomy Classification Scheme) luokittelua. Aineiston teknisiä aspekteja (laitteet, reaktiot, jne.) kuvailevaa käyttökelpoista metatietoskeemaa ei ollut valmiina. Ratkaisuna tähän päädyttiin käyttämään Yhdysvaltain kongressin kirjaston MODS-skeemaa (Metadata Object Description Schema) laajennettuna omilla "extension"-elementeillä.

Metatietojen ongelmallisimmaksi osa-alueeksi osoittautui aineiston omistajuuden ja käyttöoikeuksien määrittely. Käyttöoikeuksien osalta alalla on voimassa yleinen käytäntö, jonka mukaan mittausaineistot on vapaasti mittaukseen osallistuneiden tutkijoiden tai kollaboraation käytössä. Kirjallisia sopimuksia aineiston omistajuudesta ei ole tehty. Aineistoja ei ole aiemmin avattu tai lisensoitu mittauksen ulkopuolisille tahoille. On tiedossa, että keskeisten rahoittajien, lähinnä Euroopan Unionin ja Suomen Akatemian, uudet vaatimukset tulevat muuttamaan tätä käytäntöä tulevaisuudessa.

3.2.2 Aineiston tekniset tiedot

Pilotoitu aineisto sisältää lähinnä raakaa mittausdataa, joka on tallennettu suoraan kiihdytinlaboratorion tiedonkeruujärjestelmästä. Aineisto itsessään on tallennettu laitteistospesifisessä binääriformaatissa (nk. GREAT-formaatti), jonka määrittely on vapaasti saatavissa järjestelmätoimittajan verkkopalvelusta. Ymmärrettävyyden säilymisen varmistamiseksi tiedostomuodon kuvaus päätettiin sisällyttää siirtopakettiin.

Tämän lisäksi kokeen ajonaikainen sähköinen lokikirja, mittaussuunnitelma, mittauksen loputtua laadittu yhteenveto sekä erinäisiä mittausasetelmaa ja -laitteistoa kuvaavia dokumentteja on tallennettu PDF-muodossa. Mittalaitteistojen asetukset, esimerkiksi kalibraatioarvot, on tallennettu tekstimuodossa.

Mittausdatan käsittelyyn on julkisesti saatavilla ainoastaan kiihdytinlaboratoriossa kehitetty GRAIN-ohjelmisto. Analyysin tulokset on tallennettu AIDA XML -formaatissa, joka on alalla yleisesti käytössä oleva tiedostomuoto. Sitä tukevat useat ohjelmistot. Data-analyysissa käytetty analysointikoodi on Java-lähdekoodia. Aineiston prosessointi etenee mittalaitteen binäärimuodosta analysointikoodin avulla XML-muotoon. Säilytyksen kannalta binäärimuoto on olennaisin, sillä alkuperäinen aineisto tarvitaan, jotta aineistosta voidaan tehdä uutta tutkimusta tai verifioida aikaisempia tuloksia. XML-muoto sisältää ainoastaan etsityn fysikaalisen ilmiön siivilöitynä alkuperäisestä binäärimuodosta.

Käytetyille tiedostomuodoille ei ole juurikaan vaihtoehtoja. Binääridataa lukuun ottamatta ne ovat varsin yleisesti käytettyjä ja sopivat siten hyvin pitkäaikaissäilytykseen. Binäärimuoto on kuvailtu tutkijan näkökulmasta tarvittavalla tarkkuudella uudelleenkäyttöä varten.

3.2.3 Paketointi

Aineiston suuresta määrästä johtuen paketointiprosessi automatisoitiin ohjelmallisesti. Normaalissa mittauksessa syntyy satoja tai jopa tuhansia mittaustiedostoja, joten paketointia ei ole mielekästä toteuttaa manuaalisesti. Paketointiohjelmisto toteutettiin Python-ohjelmointikielellä käyttäen hyväksi lxml-kirjastoa, jota Tutkimuksen PAS-palvelu käyttää pakettien vastaanotossa.

Kaikki aineiston tiedostot liittyvät toisiinsa ja ne voidaan ryhmitellä tämän perusteella. Esimerkiksi osa tiedostoista koskee mittalaitetta ja sen asetuksia, osa mittalaitteen tuottamaa aineistoa ja sen tiedostomuodon kuvausta, osa aineistoa koskevia julkaisuihin ja niin edelleen. Tämä mittalaite-aineisto-tiedostomuoto-julkaisut rakenne on tärkeä aineiston ymmärrettävyyden säilymisen kannalta. Se kuvattiin siirtopaketissa rakennekarttana. Rakenne on myös hyvin yleiskäyttöinen ja sitä voidaan todennäköisesti hyödyntää myös muiden tieteenalojen tutkimusaineistoissa.

Paketoinnin alkuvaiheessa aikaa kului paljon lähinnä metatietoformaatteihin (METS, PREMIS ja MODS) tutustumiseen. Metatietokyselyssä havaittiin, että mikään yleisesti käytössä olevaa metatietoformaattia ei voi helposti käyttää kuvailemaan aineiston teknisiä metatietoja (hiukkassuihkun ja kohtiokalvojen ominaisuudet, laitteistot). Näiden kuvailuun kehitettiin oma XML-skeema, joka voidaan helposti sisällyttää MODS-metadataan käyttäen hyväksi laajennukset mahdollistavaa extension-elementtiä.

Kokonaisuudessaan aineiston paketointi KDK:n PAS-palvelun mukaisiin siirtopaketteihin onnistui lopulta varsin hyvin.

3.3 Yhteenveto

Aineistojen omistajuus- ja käyttöoikeuskysymykset osoittautuivat pilotissa haasteellisiksi. Jo pilotin alussa todettiin, että niitä ei pystytä tämän pilotin puitteissa ratkaisemaan. Asiaan pitää paneutua Kiihdytinlaboratorion sisällä, kuten myös ATT-hankkeessa.

Aineistojen laajuus asetti teknisiä vaatimuksia paketoinnin toteuttamiselle. Ratkaisuksi kehitetyn paketointityökalun käyttöönotto onnistui lopulta hyvin. Paketointityökalu on myös muiden toimijoiden hyödynnettävissä.

Tiedostomuotojen osalta mittalaitteen binäärimuodolle (GREAT-formaatti) ei ole vaihtoehtoja. Jotta aineistoja voidaan säilyttää, tulee palvelun tukea tätä tiedostomuotoa.

Kokonaisuuden kannalta olisi järkevää, jos aineiston siirto suoraan esimerkiksi Kiihdytinlaboratorion käyttämästä CSC:n IDA-palvelusta Tutkimuksen PAS-palveluun olisi mahdollista. Tällöin vältyttäisiin suurten

tietomäärien siirtelyltä useaan palveluun ja verkon kapasiteettia voitaisiin hyödyntää tehokkaasti. IDA-palvelu tarjoaisi nopeaa levypintaa aineiston käsittelyyn; PAS-palvelu puolestaan turvaisi aineiston säilyvyyden ja mahdollistaa saatavuuden tulevaisuudessa.

Pilotin puitteissa todettiin selkeä tarve selkeyttää ja monipuolistaa kuvailevan ja hallinnollisen metadatan (puoli)automaattista keräämistä jo kokeiden suorittamisen aikana. Lisäksi erittäin monimutkaisten mittauslaitteistojen ja -menetelmien kuvaamiseen tarvittavassa laajuudessa tulisi kiinnittää huomiota. Pilotissa aineiston kuvailuun kehitettiin omat ratkaisut, mutta käytäntöjen kansainvälistä kehittymistä on syytä seurata, jotta voidaan varmistua yhteentoimivuudesta kansainvälisten tahojen kanssa.

Teknisesti palvelun toiminta olisi jo nyt sillä tasolla että pilotin aineisto, tai jopa muita kiihdytinlaboratorion aineistoja, voitaisiin säilyttää Tutkimuksen PAS-palvelussa ongelmitta. Kiihdytinlaboratorio on valmis ja halukas ottamaan käyttöön palvelun myöhemmin sovittavassa laajuudessa sen siirryttyä tuotantovaiheeseen.

4 PILOTTI: TURUN YLIOPISTON AVARUUSTUTKIMUSLABORATORIO

4.1 Pilotin tiedot

Aineiston nimi	ERNE-instrumentin vuohavainnot 1996–2014 energioissa 1.6-200 MeV/n.
Aineiston koko	22 gigatavua (sisältäen mittausaineiston, mittalaitteen lähdekoodit, julkaisut ja esikatselukuvat)

Turun yliopiston fysiikan ja tähtitieteen laitoksen avaruustutkimuslaboratorio (ATL) on pienehkö tutkimusorganisaatio (yksi professori, kolme tohtoritason tutkijaa, laboratorioinsinööri ja viisi jatko-opiskelijaa), joka on mukana useissa kansainvälisissä avaruusluotainprojekteissa. Näistä vanhin, Solar and Heliospheric Observatory (SOHO), laukaistiin avaruuteen joulukuussa 1995 mukanaan ATL:n ERNE-laite, joka mittaa edelleen Auringon kiihdyttämien suurienergiaisten ionien vuota planeettainvälisessä avaruudessa. ATL:n pilottiin valittiin ERNE:n ionivuomittaukset. Koska ERNE:n havainnot muodostavat lajissaan poikkeuksellisen pitkän aikasarjan, ne ovat arvokkaita Auringon aktiivisuuden pitkäaikaisvaihteluiden ilmentäjiä. Aineiston pitkäaikaissäilytys nähtiinkin koko tiedeyhteisöä hyödyttäväksi.

Ajallisesti pilotti kesti vajaat neljä kuukautta. Pilottiin osallistuivat prof. Rami Vainio (projektin johto, aineiston valinta ja validointi, esikatselukuvien määrittely ja tuotanto), dos. Eino Valtonen (dokumentaatio, aineiston lisensointi), FM Esa Riihonen (datan tuotanto ja dokumentointi) sekä FM Timo Eronen (datan paketointi). Lisäksi datan tuotannossa tiimiä opasti raskasionivoiden osalta FM Osku Raukunen. Työaikaa pilottiin kului yhteensä noin neljä työkuukautta. Organisaation valmius aineiston pitkäaikaissäilyttämiseen oli kohtalaisen hyvä, koska datan dokumentaatio ja sitä koskeva metatieto oli suurelta osin olemassa jo ennen projektia ja koska dataa oli jo aiemmin toimitettu arkistoihin ja tietokantoihin. ATL:llä ei kuitenkaan ollut aiempaa kokemusta pitkäaikaissäilyttämiseen liittyvistä standardeista tai teknisistä määrityksistä.

4.2 Pilotin toteuttaminen

4.2.1 Metatiedot

Aineiston metatiedot selvitettiin metatietokyselyn avulla. Aineiston ymmärrettävyyden säilymisen kannalta keskeisimmät metatiedot olivat datan sisältökuvaus ja sen formaatti, instrumentin dokumentaatio, instrumenttiin ja sillä tehtyyn tieteeseen liittyvät julkaisut sekä aineiston ja sitä kuvaavien julkaisujen käyttörajoitukset ja lisensointi. Aineistoa on kerätty varsin kauan ja sen perustiedot, eli esimerkiksi

omistajuuteen ja jakeluun liittyvät tiedot, ovat varsin hyvin selvillä ja vakaassa tilassa. Aineiston vapaa jakelu on SOHO-observatorion laukaisusta vastanneiden ESA:n ja NASA:n vaatimus.

Metatiedoista haastavimmiksi osoittautuivat julkaisujen käyttöoikeusasiat. Aineisto itsessään päätettiin lisensoida Creative Commons -lisenssillä, mutta julkaisujen liittäminen aineistoon vaatii kustantajien luvat, joiden saaminen osoittautui haastavaksi. Alkuperäiset käsikirjoitukset, joita voitaisiin käyttää vapaasti, eivät ole enää kaikissa tapauksissa saatavilla. Instrumentti itsessään ja sen tuottama aineisto on merkittävässä määrin dokumentoitu näissä julkaisuissa.

Aineistoon liitettiin datatiedostojen lisäksi esikatselukuvia, joissa esitettiin vety- ja heliumionien vuo ajan funktiona kahdessa energiakanavassa kalenterivuoden mittaisilla tarkastelujaksoilla. Myös esikatselukuvien tuottamiseen alun perin suunnitellulla tavalla liittyi oikeudellisia ongelmia. Esikatseluvia tuottava työkalu oli tehty toisessa projektissa, joka rajoitti kuvien käyttöä. Lopulta kuvat päädyttiin tekemään uudelleen. Esikatselukuvien merkitys aineiston löydettävyyden ja selailtavuuden kannalta tunnistettiin keskeiseksi.

4.2.2 Aineiston tekniset tiedot

Alkuperäinen tutkimusaineisto koostuu binaarisista datatiedostoista, jotka ovat SOHO-luotaimen telemetriaa. Arkistoitava tutkimusaineisto koostuu tekstimuotoisista datatiedostoista ja PNG-muotoisista esikatselukuvista, jotka tuotetaan ATL:ssä pitkäaikaissäilytystä varten. Aineiston prosessointi alkuperäisestä telemetriasta arkistoitavaksi aineistoksi pitää sisällään myös tunnettujen fysikaalisten virheiden, kuten kuolleen ajan, korjaamisen. Nämä korjaukset tehdään ennen arkistointia, koska virheet ovat parhaiten mittalaitteen tuntevien tutkijoiden tiedossa.

Arkistoitava aineisto on valmiiksi fysikaalisissa yksiköissä ja tekstimuotoisena, joten sen käytettävyys on huomattavasti alkuperäistä aineistoa parempi, vaikka toimitus sisältääkin vain osan alkuperäisen aineiston informaatiosta. Vaihtoehto tekstimuodolle olisi ollut jokin binaarinen formaatti (esim. CDF[3] tai HDF5[4]), mutta koska tiedostojen selailtavuus olisi tuolloin menetetty, päädyttiin tekstimuotoon. Aineiston koko pysyy tästä huolimatta suhteellisen pienenä (kokonaisuudessaan 22 GB tai 1.2 GB/vuosi) ja tekstimuodossa aineistoa voi käsitellä helposti kaikilla numeerisen datan käsittelyyn tarkoitetuilla ohjelmistoilla (esim. Python, IDL® ja Matlab®) tai vaikkapa taulukkolaskentaohjelmalla.

Esikatselukuvien muoto valittiin KDK:n säilytyskelpoisten tiedostomuotojen perusteella, joten esikatselukuvienselailuun kelpaa mikä tahansa kuvankatseluohjelma.

4.2.3 Paketointi

Aineisto paketoitiin KDK:n PAS-palvelun mukaisiin siirtopaketteihin. Siirtopaketteja päätettiin tehdä kahdentyyppisiä. Yhteen kokoavaan pakettiin sisällytettiin ERNE-laitetta koskevat julkaisut ja tekniset kuvaukset. Itse aineisto sijoitettiin omiin paketteihinsa, jotka sisälsivät mittausaineiston, esikatselukuvan ja ERNE-laitteessa käytetyn on-board -ohjelmiston lähdekoodit. Näin jaoteltuna yhdestä ”pääpaketista” löytyy kaikki aineistoon liittyvät julkaisut ja aineisto on ”datapaketeissa”.

Teknisesti aineiston paketointi onnistui ilman suuria haasteita, mutta aineistossa olevien fysikaalisten suureiden kuvaaminen metatiedoissa osoittautui hankalaksi. Tämä oli erityisen olennaista siksi, että aineistosta on laskettavissa lukuisia johdannaissuureita jotka ovat merkittäviä aineiston löydettävyyden kannalta. Asiaan löydetyt tekniset ratkaisut (mm. AIDA XML) olivat teknisesti monimutkaisia ja lopulta päädyttiin vapaaseen tekstimuotoiseen kuvailuun siitä, mitä arkistoitava data on ja minkälainen suurehierarkia siihen liittyy. Tätä kuvailua täydennettiin lisäksi PNG-muotoisella kuvalla aineistoon liittyvistä johdannaissuureista. Datan paketointi onnistui määritysten mukaisesti.

[3] http://cdf.gsfc.nasa.gov/
[4] https://www.hdfgroup.org/HDF5/

4.3 Yhteenveto

Pilotin toteutuksessa oli ATL:lle joitain haasteita. Metatietokyselyssä oli muutamia hankalasti tulkittavia kysymyksiä, jotka vaikuttivat pilottiin valitun datan kannalta osittain päällekkäisiltä. Aineistoon liittyvien suureiden kuvaaminen paketissa rakennekartan muodossa osoittautui liian haastavaksi pilotin puitteissa. Sen sijaan kuvailtiin mittauksiin liittyvä suurehierarkia tekstimuodossa. Käsitteistön kehittyessä asia saattaa olla helpommin toteutettavissa. Paketoinnin rakenteen ymmärtämiseen käytettiin myös aikaa, koska aiempaa kokemusta ATL:llä aiheesta ei ollut. Toimitetun datan muokkaushistorian dokumentointi osoittautui myös jonkin verran haastavaksi, koska alkuperäistiedoston ominaisuudet eivät kopioidu suoraan valmiisiin tiedostoihin ATL:n dataputkessa, vaan alkuperäinen telemetriatiedosto pitää tunnistaa valmiin tiedoston nimen perusteella ja avata, jotta historiatietoihin pääsee käsiksi. ATL:n näkökulmasta palveluun ei kuitenkaan tarvita välttämättömiä teknisiä muutoksia, vaan aineisto voitaisiin siirtää säilytykseen pilotin jälkeen.

Pilotin hyödyt tutkimusorganisaatiolle olivat ilmeiset. ERNE-datan (ja muunkin satelliittidatan) pitkäaikaissäilytys käyttäjälle tarpeellisen metatiedon kera on välttämätöntä ratkaista joka tapauksessa, joten tämä projekti sopi organisaation tarpeisiin hyvin. Datan vapaan saatavuuden vaatimus on myös tärkeä ja siksi projektin filosofia sopi yhteen erityisen hyvin ATL:n tarpeen kanssa.

Dataa toimittavalle organisaatiolle olisi avuksi, mikäli siirtopaketin luonnista ja määrittelystä olisi tarjolla valaisevammat ja monipuolisemmat esimerkit. Todelliset paketit koostuvat useista toisiinsa liittyvistä objekteista, joten kyseisen kaltaista tilannetta kuvaava lisäesimerkki olisi tarpeen. Yksi mahdollisuus esimerkkien selkiyttämiseen olisi kommenttien lisääminen koodin sekaan. Myös mittausdatan kuvailuun paremmin sopivan käsitteistön luominen rakennekartan tekemiseen olisi tarpeen. Tämä vaatisi eri tutkimusorganisaatioiden ja arkistoivan tahon yhteistyötä, jotta asia saataisiin toteutettua mahdollisimman yleisellä tasolla.

Pilottiprojekti oli tutkimusorganisaatiolle jonkin verran työläämpi kuin ennakkoon kuviteltiin, mutta toisaalta projektin kulku oli hyvin suunniteltu ja sen seuranta ja tuki hyvin järjestetty. ATL:llä onkin valmius ja halukkuus osallistua palvelun käyttöönottoon vuoden 2016 lopussa resurssien puitteissa. Pilotissa paketointi automatisoitiin, jolloin myös myöhemmin mitattava ERNE:n ionivuodata voidaan tämän ”putken” kautta siirtää kokonaisuudessaan Tutkimuksen PAS:iin.

5 JOHTOPÄÄTÖKSET

Tutkimusaineiston elinkaari ajatellaan yleisesti laajana kokonaisuutena, joka pitää sisällään mm. tutkimussuunnitelman laatimisen, aineiston analyysin, tulosten julkaisun ja aineiston pitkäaikaissäilytyksen[5]. Pilottien käytännön kokemukset osoittavat, että säilyttämisen onnistuminen edellyttää asian huomioimista koko aineiston elinkaaren ajan. Pilottien tavoitteena oli selvittää mitä vaatimuksia aineistojen ymmärrettävyyden säilyttäminen asettaa Tutkimuksen PAS-palvelulle. Seuraavassa käsitellään pilottien keskeisiä havaintoja Tutkimuksen PAS-palvelun vaatimusmäärittelyä ajatellen aineistojen elinkaarta mukailevassa järjestyksessä.

Tutkimussuunnitelma. Aineistojen säilyttämiskelpoisuus oli määritelty merkittävissä määrin jo ennen varsinaisen tutkimuksen alkamista tutkimuksen suunnitteluvaiheessa. Käytännössä tämä tarkoitti esimerkiksi mittaussuunnitelmaa (Kiihdytinlaboratorio), eettisen toimikunnan lupapyyntöä (BML) tai aineiston jakelusta tehtyä sopimusta (ATL). Kaikissa näissä tapauksissa aineiston elinkaaren aivan

[5] Borg, Sami & Kuula, Arja (2007). Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari. Valmisteluraportti OECD:n datasuosituksen toimeenpanomahdollisuuksista Suomessa. Tampere: Tampereen yliopisto. Yhteiskuntatieteellisen tietoarkiston julkaisuja; 6.

alkuvaiheissa, käytännössä jo tutkimussuunnitelmassa, tehdyt valinnat vaikuttivat siihen kuinka aineistoa voidaan uudelleen käyttää alkuperäisen tutkimuksen jälkeen. Aineisto voi olla täysin avointa (ATL), sensitiivistä (BML) tai täysin luvanvaraista (Kiihdytinlaboratorio). Erityisesti aineistojen jakelu kolmansille osapuolille (saatavuus) edellyttää käyttöoikeuksien täsmällistä määrittelyä. *Tutkimuksen PAS-palvelun tulisi tukea tutkimusaineistojen käyttöoikeuksien ja rajoitusten kuvaamista.*

Aineiston käsittely. Aineistojen käsittely parantaa aineiston hyödynnettävyyttä tulevaisuudessa. Piloteissa aineistolle tehtiin luonteeltaan kahdenlaista käsittelyä: siivilöintiä (Kiihdytinlaboratorio) ja korjausta (ATL). Siivilöinnissä alkuperäisestä aineistosta poistetaan tarpeettomat osat ja korjauksessa puolestaan korjataan aineistoon päätyneitä tunnettuja virheitä. On täysin tapauskohtaista, mikä käsittelyvaihe soveltuu aineiston jatkokäytön kannalta parhaiten säilytettäväksi. Myös raakadatan säilyttäminen voi olla perusteltua (BML). Parhaiten asian osaavat arvioida aineiston tuottaneet tutkijat itse. Säilyttäminen aineiston kaikissa vaiheissa edellyttää tukea tarvittaville tiedostomuodoille ja tarvittavaa kapasiteettia. *Tutkimuksen PAS-palvelun tulisi mahdollistaa tutkimusaineiston säilyttäminen sen käsittelyn kaikissa vaiheissa.*

Tulosten julkaisu. Tutkimustulosten tieteellinen julkaisut voi olla merkittävä myös aineiston jatkokäyttöä ajatellen ja halutaan säilyttää. Esimerkiksi aineiston tai siihen liittyvän mittalaitteen keskeinen dokumentaatio (ATL) voi olla tieteellisessä julkaisussa. Kustantajan kanssa tehty julkaisusopimus voi estää tai rajoittaa julkaisujen säilyttämistä. Julkaisun alkuperäisen käsikirjoituksen tekijänoikeudet ovat kuitenkin tutkijalla itsellään, joten niiden säilyttäminen mahdollisimman aikaisessa vaiheessa tulisi mahdollistaa. Jos julkaisut eivät sisällä säilyttämisen kannalta välttämätöntä tietoa, pitäisi niihin kuitenkin pystyä viittaamaan säilytettävästä aineistosta. *Julkaisujen säilyttäminen tai niihin viittaaminen tulisi olla mahdollista Tutkimuksen PAS-palvelussa.*

Aineiston pitkäaikaissäilyttäminen. Pilotoitujen aineistojen tiedostomuodot käytiin läpi ja tunnistettiin niiden vaihtoehdot, joskin todellisia vaihtoehtoja oli melko vähän. Kaikki aineistot olivat laskennallisia ja niihin liittyi jokin mittalaite, jolla ne olivat tuotettu. Tämä tarkoitti sitä, että aineisto oli alkuperäisesti mittalaitteen tuottamassa ”custom-formaatissa”, joka oli joko mittalaitteen (BML) tai tutkijoiden itsensä kehittämä (Kiihdytinlaboratorio ja ATL). Ymmärrettävyyden säilyttäminen tässä tapauksessa on haastavaa ja jotta se ylipäätänsä olisi mahdollista, täytyy formaatille ja sen dokumentaatiolla luoda kriteeristö. *Tutkimuksen PAS-palvelussa tulisi olla ennalta määritetyt kriteerit tutkija- tai laitekohtaisten tiedostomuotojen vastaanottamiselle.*

Tutkija- tai laitekohtaisesta tiedostomuodosta aineisto voi olla mahdollista käsitellä johonkin muuhun tiedostomuotoon, mutta sen käyttökelpoisuus on tapauskohtaista. BML:n ja Kiihdytinlaboratorion tapauksessa tiedostomuodot olivat binäärisiä eikä niille ollut vaihtoehtoja ilman, että niistä olisi menetetty informaatiota. ATL:n tapauksessa aineisto oli tekstimuotoinen, mutta ei noudattanut mitään yleisesti käytettyä tekstimuotoisen tiedon esitystapaa (esim. CSV tai TSV, Comma/Tabular Separated Value). Ymmärrettävyyden säilymisen kannalta tekstimuotoisen tiedon tallentamisessa tulisi kannustaa yleisten esitystapojen käyttöön ja tapauskohtaisesti voi olla perusteltua normalisoida tekstimuotoisia aineistoja näihin muotoihin. *Tutkimuksen PAS-palvelun tulisi tukea yleisesti käytettyjä tekstimuotoisen tiedon esitystapoja.*

Aineistojen tiedostomuotoon voi liittyä myös pakkausalgoritmi (BML). Säilyttämisen kannalta pakkausalgoritmit ovat ongelmallisia, koska koko aineiston ymmärrettävyys riippuu tästä algoritmista. Pakkaus voi pienentää aineiston vaatimaa kapasiteettitarvetta, mutta yleisesti tämä on tarpeetonta, sillä bittitason säilytysratkaisun tarjoava KDK:n PAS-palvelun kanssa yhteinen PAS-ratkaisu pakkaa sisäisesti kaikki aineistot. Kuitenkin on poikkeustapauksia, joissa tiedostomuotoon itseensä kuuluu pakkaus, ja tämän pakkauksen poistaminen heikentäisi aineiston ymmärrettävyyttä tulevaisuudessa. *Tästä syystä Tutkimuksen PAS-palvelun tulisi tapauskohtaisesti tukea myös pakattuja tiedostomuotoja.*

Jokaiseen aineistoon liittyi myös yleisesti käytettyjä tiedostomuotoja (PDF, PNG), jotka kuvaavat tai täydentävät varsinaista aineistoa (dokumentaatio, esikatselukuvat). *Tutkimuksen PAS-palvelun tulisi tukea yleisesti käytettyjä teksti-, kuva- ja videotiedostomuotoja.*

Aineistojen metatiedot selvitettiin metatietokyselyn (Liite A) avulla ja lähtökohtana oli selvittää kaikki tutkimusaineistoon liittyvät metatiedot ja arvioida näiden merkitys säilytyksen kannalta. Olennaisiksi metatiedoiksi säilytyksen kannalta tunnistettiin sekä kaikille tieteenaloille yhteisiä kuvailevia metatietoja (aineiston perustiedot, tieteenalaluokitus, asiasanat) että jokaiselle tieteenalalle ominaista metatietoja, kuten muuttujatiedot. Tällaisia muuttujatietoja olivat esimerkiksi törmäysenergiat (Kiihdytinlaboratorio), mittausresoluutio (BML) ja hiukkasten energiat (ATL), jotka ovat keskeisiä aineiston löydettävyydelle tulevaisuudessa. *Tutkimuksen PAS-palvelun tulee tukea sekä tutkimusaineistojen yhteisiä kuvailevia metatietoja (aineiston perustiedot) että tieteenalakohtaisia kuvailevia metatietoja (esimerkiksi muuttujatiedot).*

Pilotoiduista aineistoista tunnistettiin hyvin yleistettäviä rakenteita, joiden mukaan aineistoihin liittyvät tiedostot voidaan ryhmitellä. Tällaisia rakenteita olivat esimerkiksi: aineisto, mittalaite, mittalaitteen kuvaus, tiedostomuodon kuvaus, mittauslupa, julkaisut (ATL ja Kiihdytinlaboratorio). Näiden rakenteiden kuvaaminen kaikissa aineistoissa samalla tavalla edistää niiden ymmärrettävyyden säilymistä, joten nämä rakenteet kuvattiin myös aineistojen metatiedoissa. Toisaalta on myös olemassa tieteenalakohtaisia rakenteita (BML), eli alan aineistot on tapana lähes aina ryhmitelty määrätyllä tavalla. *Tutkimuksen PAS-palvelun tulisi mahdollistaa aineistoihin liittyvien yleisten rakenteiden kuvaamisen konsistentilla tavalla.*

Aineiston paketointi teknisesti ei ollut vaativaa ja KDK:n paketointimääritykset osoittautuivat toimiviksi myös tutkimusaineistoille. Suurien (Kiihdytinlaboratorio) ja jatkuvasti kasvavien (ATL) aineistojen paketointi ja siirto säilytykseen on järkevää automatisoida ja integroida osaksi nykyistä aineiston tuotantoprosessia. *Tätä varten Tutkimuksen PAS-palvelun tulisi tarjota yleinen paketointikomponentti, joka mahdollistaisi siirtopakettien luomisen automatisoinnin.*

Kokonaisuudessaan piloteista saatiin kerättyä merkittävä määrä hyödyllistä tietoa vaatimusmäärittelyn tueksi sekä paljon hyödyllistä tietoa tutkijoiden aineistoista ja käytännön työhön liittyvistä ongelmista. Samoin piloteissa onnistuttiin parantamaan osallistuneiden organisaatioiden teknisiä valmiuksia tulevan palvelun käyttöön.

LIITE A: METATIETOKYSELY

Metatietokysely pilotin 1. tarkastuspisteeseen.

1. Kuka on tehnyt tutkimusaineiston?
   [esim. tekijä, author]
2. Kuka omistaa tutkimusaineiston?
   [esim. omistaja, owner]
3. Kuka antaa tutkimusaineiston käyttöön?
   [esim. jakelija, distributor]
4. Miten tutkimusaineisto on kerätty?
   [esim. menetelmän kuvaus, method]
5. Miten menetelmästä saa lisätietoa?
   [esim. menetelmään liittyvät julkaisut, linkit]
6. Miten tutkimusaineistoa saa käyttää?
   [esim. käyttöehdot, rights]
7. Miten aineistosta saa lisätietoa?
   [esim. yhteystiedot, distributor contact]
8. Miten aineistoon linkitetään?
   [esim. aineiston tunniste, identifier]
9. Miten tutkimusaineistoon viitataan julkaisussa?
   [esim. lähdeviitteen muoto, citation]
10. Miksi tutkimusaineisto on tuotettu ja avattu?
    [esim. tutkimusaineiston hallintasuunnitelma, DMP]
11. Mihin tutkimusaineisto on tarkoitettu?
    [esim. aihe ja kuvaus, subject, description]
12. Mikä on tutkimusaineiston nimi?
    [esim. aineiston nimi, title]
13. Mitä tutkimusaineisto sisältää?
    [esim. kuvaus, description]
14. Mihin tutkimusalaan aineisto liittyy?
    [esim. tieteenalaluokitus, classification]
15. Mihin asioihin tutkimusaineisto liittyy?
    [esim. aihe, subject]
16. Minkä kielinen aineisto on?
    [esim. aineiston kieli, language]
17. Mitä muuttujatietoa aineistossa on käytetty?
    [esim. muuttujatieto, dimension]

18. Mihin julkaisuun aineisto liittyy?
    [esim. julkaisu, publication]
19. Missä organisaatiossa aineisto on tehty?
    [esim. organisaatio, organization]
20. Missä projektissa aineisto on tehty?
    [esim. projekti, project]
21. Missä maassa aineisto on tehty?
    [esim. maa, country]
22. Missä aineisto on löydettävissä?
    [esim. katalogi, catalog]
23. Missä muodossa aineisto on saatavilla?
    [esim. tiedostomuoto, file format]
24. Miltä ajalta tutkimusaineisto on olemassa?
    [esim. ajallinen kattavuus, temporal]
25. Milloin tutkimusaineisto on avattu?
    [esim. avaushetki, issued]
26. Milloin tutkimusaineistoa on muokattu?
    [esim. muokkaushetki, modified]