AVOIN TIEDE
JA TUTKIMUS

TUTKIMUSDATAN PITKÄAIKAISSÄILYTYS: KANSAINVÄLINEN KATSAUS

Julkaisu Tutkimusdatan pitkäaikaissäilytys: Kansanvälinen katsaus	
Julkaisija Avoin tiede ja tutkimus -hanke	**Julkaisuajankohta** 2.11.2015
Tekijä Tutkimus-PAS -työryhmä	
Lisenssi	
Julkaisun jakelu PDF-tiedosto ladattavissa avointiede.fi/keskeiset-julkaisut http://urn.fi/URN:NBN:fi-fe2016122731712	
Yhteystiedot http://avointiede.fi avointiede@postit.csc.fi	

SISÄLLYS

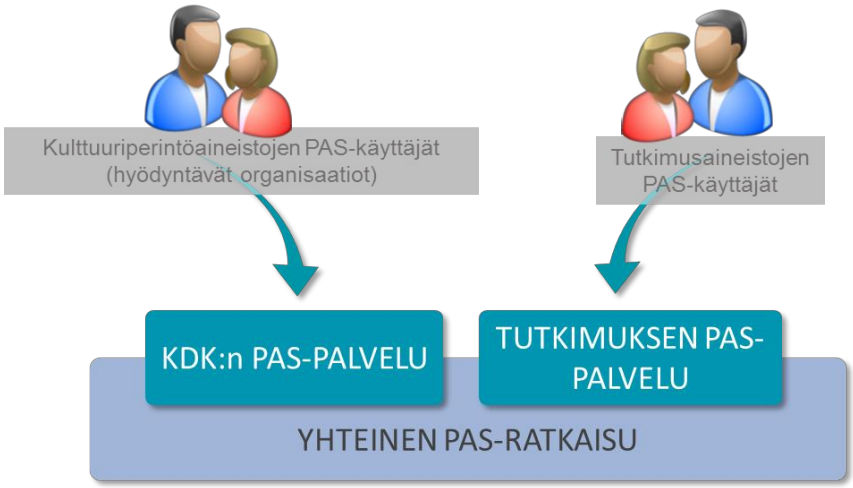
1 JOHDANTO

Opetus- ja kulttuuriministeriö on käynnistänyt tiedon saatavuuden ja avoimen tieteen edistämiseksi Avoimen tieteen ja tutkimuksen hankkeen (ATT) vuosille 2014–2017. Tavoitteena on, että vuoteen 2017 mennessä Suomi nousee yhdeksi johtavista maista tieteen ja tutkimuksen avoimuudessa ja 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.

Tutkimuksen pitkäaikaisen saatavuuden varmistaminen on yksi ATT-hankkeen tavoitteista ja kansainvälisen tilannekuvan muodostaminen on vuosille 2015–2017 tehdyssä etenemissuunnitelmaehdotuksessa tunnistettu yhdeksi tutkimuksen PAS-palvelun toteutuksessa tarpeelliseksi pohjatyöksi. Tässä dokumentissa kuvataan kansainvälisen tilannekuvan kartoituksen tulokset.

Pitkäaikaissäilytys tarkoittaa digitaalisen informaation säilyttämistä ymmärrettävänä ja käytettävänä useiden kymmenien ja jopa satojen vuosien ajan. Laitteet, ohjelmistot ja tiedostomuodot vanhenevat ajan myötä, mutta informaation täytyy säilyä. Luotettava pitkäaikaissäilyttäminen ja siten aineistojen pitkäaikainen saatavuus edellyttää sisällön eheyden aktiivista valvontaa ja monenlaisiin riskeihin varautumista. Tässä ovat keskeisessä asemassa metatiedot, jotka kuvailevat mm. aineiston sisällön, historian ja alkuperän sekä tiedot siitä, miten informaatiota voidaan käyttää.

Tässä tilannekartoituksessa keskityttiin erityisesti palveluratkaisuihin ja prosesseihin sekä niitä tukeviin yhteistyön ja hallinnon järjestelyihin. Kansallisen digitaalisen kirjaston (KDK) PAS-ratkaisun toteutuksessa on jo tutustuttu tietojärjestelmä- ja teknologiatason tilanteeseen ja kehitystrendeihin. KDK:n PAS-ratkaisu on tarkoituksena ottaa sellaisenaan myös tutkimuksen PAS-palvelun pohjaksi, joten tietojärjestelmä- ja teknologiatason kysymyksiin ei tässä tilannekartoituksessa ollut tarpeen ottaa kantaa. Erityisesti bittitason säilytyksessä tukeudutaan täysin KDK:n PAS-palvelun yhteydessä toteutettuun PAS-ratkaisuun ja säilytyksessä käytetään samoja ohjelmistoja, laitteistoja ja medioita (ks. kuva 1). Edelleen kartoituksessa pyrittiin keskittymään kohteisiin, joilla on vastattavanaan samantapaisia erityiskysymyksiä kuin ATT-hankkeessa. Ne liittyvät erityisesti tieteenalojen ja organisaatioiden moninaisuuteen ja siitä seuraaviin yhteentoimivuuden ja yhdenmukaisuuden haasteisiin.



Kuva 1: KDK:n ja tutkimuksen PAS-palvelut

Näistä lähtökohdista päätettiin selvittää neljän organisaation ratkaisut erityisesti yllä mainittuihin kysymyksiin. Organisaatioiksi valittiin

- CINES (Ranska),
- DANS (Alankomaat),
- NorStore (Norja) ja
- UKDA (Iso-Britannia).

Nämä organisaatiot valittiin mukaan kartoitukseen, koska ne toimivat aktiivisesti digitaalisen pitkäaikaissäilytyksen parissa ja niihin oli valmiina luotettavat ja kiinteät suhteet pitkäaikaissälytykseen liittyvien hankkeiden kautta. Selvitys toteutettiin Webropol-kyselyllä kesä-heinäkuussa 2015 (ks. liite B). Ennen kyselyn käynnistämistä organisaatioihin otettiin yhteyttä, niille kerrottiin kyselyn tavoitteet ja varmistettiin halu osallistua kyselyyn.

Tässä raportissa tehdään yhteenveto organisaatioiden tarjoamista säilytyspalveluista kyselyvastauksiin ja palveluiden julkisiin dokumentaatioihin perustuen. Lisäksi organisaatioiden toimintamalleja ja prosesseja on soveltuvin osin verrattu jo tuotannossa olevan KDK:n PAS-palvelun prosesseihin tai toteutukseen.

2 TAUSTATIEDOT

Kuten yllä todettiin, tilannekuvan kartoittamiseksi selvitettiin neljän organisaation (CINES, DANS, NorStore ja UKDA) pitkäaikaissäilytyspalveluiden toimintamallit ja prosessit. Seuraavassa on kuvattu nämä organisaatiot ja niiden tarjoamat palvelut yleisellä tasolla.

CINES (Centre Informatique National de l'Enseignement Supérieur)[1] on Ranskan korkeakoulu- ja tutkimusministeriön alainen laitos, joka tuottaa suurlaskennan, digitaalisen pitkäaikaissäilytyksen sekä tietojärjestelmien ylläpidon palveluita. CINES vastasi kyselyyn PAC (Plateforme d'Archivage du Cines)[2] palvelun osalta. CINES säilyttää digitaaliset versiot väitöskirjoista kansallisen lainsäädännön nojalla. Tutkimusaineistojen säilyttämiseen CINES on saanut toimeksiannon Ranskan korkeakoulutus- ja tutkimusministeriöltä (Ministère de l'Enseignement supérieur et de la Recherche, MESR). CINES tekee yhteistyötä muiden kansallisten palveluiden sekä useiden kansanvälisten projektien kanssa (esimerkiksi EUDAT[3], Open Preservation Foundation[4], Research Data Alliance[5]).

DANS (Data Archiving and Networked Services)[6] tarjoaa aineiston arkistoinnin ja uudelleenkäytön palveluita sekä niihin liittyviä koulutus- ja konsultointipalveluita. DANS on Alankomaiden kuninkaallisen tiedeakatemian ja NOW:n (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) alainen laitos. DANS vastasi kyselyyn EASY[7] palvelun osalta. DANSin tarjoamat muut palvelut ovat (1) Research information system (NARCIS)[8], (2) Virtual Research Environment (Dutch Dataverse)[9] ja (3) Persistent identifier resolver[10]. DANS tekee yhteistyötä kansainvälisten infrastruktuuriprojektien kanssa (esim. DARIAH[11], EUDAT) ja DANSia vastaavien instituuttien kanssa (esim. UKDA, GESIS[12]).

NorStore[13] on Norjan digitaalisen tieteellisen datan hallinnan ja pitkäaikaisarkistoinnin kansallinen infrastruktuuri, joka tarjoaa tallennuksen ja tiedonsiirron resursseja. NorStore vastasi kyselyyn NorStore Research Data Archiven palvelun kannalta[14]. Koska NorStoren palvelu on vasta pilottivaiheessa, heillä ei ollut vastauksia yksityiskohtaisiin, erityisesti teknisiin kysymysiin. NorStore aikoo myöhemmin integroida palvelunsa EUDATin palveluihin.

UKDA (UK Data Archive)[15] kerää talteen, hoitaa ja asettaa käyttöön yhteiskuntatieteellisen ja humanistisen alan tutkimusdataa. UKDA on osa Essexin yliopistoa, ja se on pitänyt yllä tutkimusdatapalveluita vuodesta 1967. UKDA on CESSDA:n (Consortium of European Social Science Data Archives)[16] jäsen. Se tekee

[1] https://www.cines.fr/en/
[2] https://www.cines.fr/en/long-term-preservation/
[3] http://eudat.eu/
[4] http://openpreservation.org/
[5] https://rd-alliance.org/
[6] http://www.dans.knaw.nl/en
[7] http://easy.dans.knaw.nl
[8] http://www.narcis.nl/
[9] http://www.dataverse.nl
[10] http://persistent-identifier.nl/
[11] https://www.dariah.eu/
[12] http://www.gesis.org/
[13] https://www.norstore.no/
[14] http://archive.norstore.no
[15] http://www.data-archive.ac.uk
[16] http://cessda.net/

yhteistyötä muiden vastaavien palveluiden, erityisesti ICPSR:n (Inter-university Consortium for Political and Social Research)[17], kanssa.

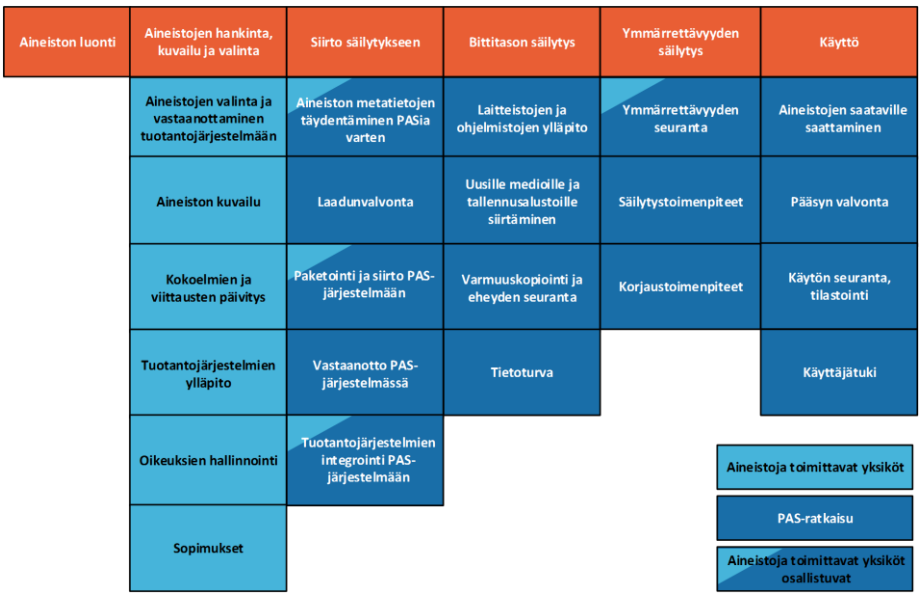
Kaikki kyselyyn osallistuneet organisaatiot säilyttävät raakadataa ja käsiteltyä dataa ja UKDAta lukuun ottamatta myös julkaisuja. CINES ja NorStore hyväksyvät aineistoja kaikilta tutkimusaloilta, kun taas DANS ja UKDA keskittyvät humanistisiin ja yhteiskuntatieteellisiin tutkimusaloihin. Aineistoja kaikissa kyselyyn osallistuneissa palveluissa on melko vähän verrattuna lukuihin, joita käytettiin tutkimuksen PAS-palvelun kustannus- ja hyötyanalyysissä; UKDA 2.5 Tb, DANS ja NorStore 50 Tb, ja CINES 100 Tb. Kyselyssä pyrittiin myös selvittämään, miten organisaatiot ennustavat datamäärän kasvavan vuoteen 2020 mennessä, mutta organisaatiot eivät osanneet tätä kovinkaan tarkasti arvioida. UKDAn mielestä datamäärän kasvua on mahdotonta ennakoida.

Kapasiteetin hankinta ja siihen liittyvät muut kustannukset ovat keskeinen tekijä palvelun kustannuksia arvioitaessa (vrt. tutkimuksen PAS-palvelun kustannus- ja hyötyanalyysi). Todellinen kapasiteettitarve riippuu aineistomäärän lisäksi muun muassa säilytettävän aineiston valmiusasteesta, eli siitä, kuinka pitkälle aineisto on prosessoitu (raakadata vs. käsitelty data) sekä siitä, millaisia tiedostomuotoja käytetään (pakatut binääriformaatit vs. tekstitiedostot).

Valituille organisaatioille kohdennettu kysely jaettiin viiteen osioon:

- aineistojen hankinta,
- siirto säilytykseen,
- bittitason säilytys,
- ymmärrettävyyden säilyttäminen ja
- käyttö.

Jaottelu perustuu löyhästi LIFE-elinkaarimalliin (ks. kuva 2), jota käytettiin myös tutkimuksen PAS-palvelun kustannus- ja hyötyanalyysissä. Seuraavissa luvuissa on vedetty yhteen organisaatioiden vastauksia tämän jaottelun mukaisesti.



Kuva 2: LIFE-elinkaarimallin vaiheet

[17] https://www.icpsr.umich.edu/icpsrweb/landing.jsp

3 AINEISTOJEN HANKINTA

Aineistojen hankinta -osiossa selvitettiin

- ketkä ovat palveluiden asiakkaita ja kuka heidät valitsee,
- kuka valitsee säilytettävän aineiston ja millä perusteella,
- kuka omistaa säilytyksessä olevan aineiston,
- miten palvelu on rahoitettu ja
- millaisia lisenssejä säilytykseen otettavalta aineistolta vaaditaan.

Aineistojen hankinnassa erot KDK:n PAS-palveluun ovat merkittävät. KDK:n PAS-palvelussa säilytettävien aineistojen hankinta ja valinta perustuu pitkälti hyödyntävien organisaatioiden lakisääteisiin velvoitteisiin. Tutkimuksen PAS-palvelun osalta nämä asiat on sovittava muilla tavoin.

CINESin palveluun aineistoa tallettavat[18] vain organisaatiot (depositor services), jotka myös valitsevat säilyttäjät ja säilytettävän aineiston. Aineistojen valintaperusteista päättävät näiden organisaatioiden tieteelliset komiteat. Instituutioiden tekninen henkilökunta, yhdessä CINESin asiantuntijoiden kanssa, on vastuussa aineiston teknisistä piirteistä.

DANSin palveluun aineistoa tallettavat tutkijat itse, mutta DANS tarkistaa kaikki talletukset. DANSissa säilytettävää aineistoa ei kukaan erityisesti valikoi; ainoastaan aineiston tekniset piirteet arvioidaan. NorStoren palveluun aineistoa tallettavat tutkijat sillä edellytyksellä, että tutkimus on (1) julkisesti rahoitettua, (2) aineisto ei sisällä sensitiivistä dataa ja (3) aineiston omistaja tarjoaa aineiston avoimesti saataville.

UKDAn palveluun aineistoa tallettavat tutkijat; erityisesti ne, joita ESRC (Economic and Social Research Council)[19] tai valtion laitokset tukevat taloudellisesti. UKDAlla on varsin tarkat ja hyvin dokumentoidut valinta- ja arviointikriteerit[20,21], joilla pyritään varmistamaan, että säilytykseen valitaan vain olennaista ja vaikuttavaa dataa. Tällä pyritään varmistamaan, ettei resursseja käytetä aineistoihin, joilla ei ole pitkäaikaista arvoa. Aineiston tallettajan tulee hakea säilytysmahdollisuutta UKDAlta. Hakemuslomakkeessa aineiston tallettajan tulee vastata joukkoon kysymyksiä ja niiden kautta perustella aineiston säilytystarve. UKDAn Data Appraisal Group (DAG) arvioi hakemukset ja luokittelee säilytyskategorian. Käytössä on neljä kategoriaa: (1) aineisto valitaan pitkäaikaissäilytykseen; (2) aineisto valitaan lyhytaikaiseen säilytykseen (bittitason säilytys); (3) aineisto tarjotaan saataville UKDAn palveluiden kautta; tai (4) aineiston metatiedot haravoidaan UKDAn kautta saataville. Aineistot, jotka aluksi luokitellaan kategorioihin 2 tai 3, voidaan myöhemmin siirtää kategoriaan 1.

UKDA painottaa aineistojen valinnassa seuraavia seikkoja: (1) aineiston merkityksellisyys valitun strategian valossa; (2) tieteellinen tai historiallinen arvo; (3) uudentyyppisyys; (4) kansainvälinen arvo, (5) ainutlaatuisuus tai katoamisriski; (6) käyttökelpoisuus; ja (7) toistettavuuden tarve. Aineistoa ei kuitenkaan koskaan oteta säilytykseen, jos aineistoon liittyy juridisia tai eettisiä ongelmia, jos metatiedot eivät mahdollista uudelleenkäyttöä tai jos tiedostomuodot eivät mahdollista uudelleenkäyttöä.

DANSin ja UKDAn palvelussa aineistot omistavat tallettaja ja rahoittaja, CINESin palvelussa rahoittaja. Palvelujen rahoituslähteet on esitetty taulukossa 1.

[18] Tässä dokumentissa tallettajalla tarkoitetaan henkilöä tai järjestelmää, joka siirtää säilytettävän tiedon PAS-palveluun. OAIS-viitemalli käyttää tästä roolista termiä ”tiedon tuottaja” (producer).
[19] http://www.esrc.ac.uk/
[20] http://ukdataservice.ac.uk/media/455175/cd234-collections-appraisal.pdf
[21] http://ukdataservice.ac.uk/media/398725/cd227-collectionsdevelopmentpolicy.pdf

Taulukko 1: Palveluiden rahoituslähteet

	Vuosittainen julkinen rahoitus	Kertakorvaus julkinen rahoitus	Rahoitus kaupallisilta toimijoilta	Rahoitus tallettajilta	Rahoitus voittoa tavoittelemattomilta rahoittajilta	Rahoitus julkisilta rahoittajilta	Rahoitus myynnin avulla	Lyhytaikainen projektiperustainen rahoitus
CINES	X			X		X		
DANS		X		X				X
NorStore[22]	X					X		
UKDA[23]		X				X		X

Kaikilla palveluilla on kannustimia aineiston säilyttämiseen. CINESissä kannustimena toimivat kansalliset suositukset, DANSissa se, että rahoittajat yhä useammin vaativat datan hallintasuunnitelman (Data Management Plan). NorStore tarjoaa ilmaisen säilytyksen kymmeneksi vuodeksi, ja UKDAn kohdalla säilyttäminen on vaatimus ESRC:n rahoittamalle tutkimukselle.

Aineistoilta vaadittavat ja suositeltavat lisenssit on esitetty taulukossa 2.

Taulukko 2: Palveluiden säilytykseen otettavien aineistojen lisenssit (P=pakollinen; S=suositeltava)

	Copyright protected, no licenses	International licenses	National licenses	Public domain	Open licenses (e.g. Creative Commons)	Uniform collection of licenses of various types
CINES	P		P	P	P	
DANS	P	P	P	P	P, S	P
NorStore			P, S		P, S	
UKDA		P, S	P, S	P, S	P, S	P, S

[22] NorStoren rahoituksesta vastaa kansallinen tutkimus neuvosto (1/3) ja yliopistot (2/3)
[23] UKDA:n rahoituksesta vastaa ESRC ja projektit joihin UKDA osallistuu

4 SIIRTO SÄILYTYKSEEN

Siirto säilytykseen -osiossa selvitettiin

- teknisiä vaatimuksia säilytettävälle aineistolle sekä sitä,
- miten varmistetaan, että teknisiä vaatimuksia noudatetaan.

Teknisten vaatimuksien osalta pyrimme erityisesti selvittämään palveluiden asettamia vaatimuksia aineiston

- paketoinnille,
- kuvailevalle metatiedolle,
- historia- ja alkuperämetatiedolle,
- tekniselle metatiedolle,
- käyttöoikeustiedolle ja
- rakenteelliselle metatiedolle.

CINES on kehittänyt oman metatietoformaatin aineiston siirtämiseksi palveluun. Sitä on käytettävä kaiken tyyppisen aineiston sisällönkuvailussa[24], ja sitä käytetään myös kaiken muun metatiedon esittämiseen. Myös DANSissa paketointi on pakollista; aineisto on dokumentoitava käyttäen Dublin Core -metatietoformaattia. Sen lisäksi DANSissa voidaan käyttää muita, eri tutkimusaloille ominaisia metatietoformaatteja. UKDA ei vaadi erityistä paketointia, sillä aineisto valmistellaan säilytykseen paljolti vasta UKDAssa. Taulukossa 3 on esitetty kyselyyn osallistuneiden palveluiden vaatimukset erilaisille metatiedoille.

Taulukko 3: Metatieto vaatimukset (P=pakollinen; S=suositeltava; V=vapaaehtoinen)

	Paketointi	Kuvaileva metatieto	Historia- ja alkuperätieto	Tekninen metatieto	Käyttöoikeustieto	Rakenteellinen metatieto
CINES[25]	P	P	P	P	P	P
DANS	P	P	S	P	P	
NorStore	S	P	V		P	
UKDA		P	P	P	P	P
KDK-PAS	P	P	P	P	S	P

NorStore ei ole määritellyt säilytykseen otettavia tiedostomuotoja, mutta muilla organisaatioilla on lista sallituista tiedostomuodoista. CINES on määritellyt[26] 20 tiedostomuotoa, jotka ovat säilytyskelpoisia

[24] http://www.cines.fr/pac/sip.xsd
[25] CINESin vaatimukset perustuvat paketoinnissa käytettäviin skeemoihin
[26] https://www.cines.fr/en/long-term-preservation/expertises/formats-expertise/archiving-format-list/

(*archivable formats*)[27]. KDK:n, DANSin ja UKDAn ratkaisusta poiketen CINES ei ole erikseen määritellyt siirtokelpoisia tiedostomuotoja.

DANS on määritellyt[28] 26 säilytyskelpoista tiedostomuotoa (*preferred formats*) ja 23 siirtokelpoisia tiedostomuotoa (*acceptable formats*)[29]. DANSissa on työryhmä, joka seuraa tiedostomuotojen kehittymistä. Toisin kuin esimerkiksi KDK:n PAS-palvelussa, DANS ei muunna siirtokelpoisia tiedostomuotoja säilytyskelpoisiksi vaan säilyttää ne sellaisenaan.

UKDA on määritellyt[30] yli 20 tiedostomuotoa säilytyskelpoisiksi (*recommended formats*) ja yli 30 tiedostomuotoa siirtokelpoisiksi (*acceptable formats*). UKDAlla on vuosittainen tiedostomuotojen katselmus, joka perustuu kohdeyhteisön tarpeisiin ja tiedostomuotojen riskeihin.

Liitteessä A on listattu palveluiden hyväksymät säilytys- ja siirtokelpoiset tiedostomuodot. Organisaatioiden julkaisemien tiedostomuotoluetteloiden taso vaihtelee jonkin verran. CINES kertoo listauksessaan myös sallitut versiot ja niiden PRONOM-tunnisteet[31]; näin on tehty myös KDK:ssa. DANS ei ota kantaa tiedostomuotojen versioihin. UKDA listaa vain esimerkkejä soveltuvista tiedostomuodoista, eikä ota kantaa tiedostomuotojen versioihin. Tämä ratkaisu toimii UKDAssa, koska, kuten on todettu, UKDA valmistelee tutkimusdatan säilytykseen itse ja koska datamäärät ovat varsin maltillisia. Esimerkiksi KDK:n PAS-palvelun tapauksessa tällainen ratkaisu ei toimisi, se pyrkii mahdollisimman pitkälle automatisoituun prosessiin.

CINESissä määritysten noudattamisen varmistaminen tapahtuu varsin samalla tavalla kuin KDK:n PAS-palvelun käyttöönottoprosessissa[32]. Valmisteluvaiheessa (*preparation phase*) aineiston laatutaso määritellään yhdessä aineiston tuottajien kanssa. Aluksi muutamia siirtopaketteja tehdään käsin sen varmistamiseksi, että osapuolet ymmärtävät toisiaan. Nämä siirtopaketit siirretään vastaanoton testiympäristöön. Tämän jälkeen prosessi automatisoidaan ja siirrytään tuotantoympäristöön. Ennen tuotantoon siirtymistä CINESin henkilökunta tekee vielä testejä, joilla varmistetaan koko tuotantoketjun toiminta. Tuotantovaiheessa kaikki toimii automaattisesti: Tiedostomuotojen oikeellisuus tarkistetaan erilaisten työkalujen (kuten Jhove[33]) avulla, siirtopaketin rakenne tarkistetaan XSD-validoinnilla jne. Jos automaattinen aineiston tarkastaminen epäonnistuu, aineisto hylätään. Joissakin harvinaisissa tapauksissa, kun aineiston tuottaja ei pysty toimittamaan riittävän hyvälaatuista aineistoa, sopimusta CINESin ja aineistoa toimittavan instituution välillä voidaan muuttaa.

DANSissa vastaanotto on enemmän henkilötyötä vaativaa, sillä aineiston dokumentaatio tarkistetaan manuaalisesti (*checked by data archivist*). Jos aineisto tai sen dokumentaatio ei noudata sille asetettuja

[27] Säilytyskelpoinen tiedostomuoto on sellainen, jossa PAS-palvelu säilyttää digitaalisia objekteja. Tietosisällön säilyminen ja ymmärrettävyys voidaan taata, koska tiedostomuodon migraatioon (tai emulointiin) on tekniset valmiudet.

[28] http://www.dans.knaw.nl/en/deposit/information-about-depositing-data/dans-preferred-formats-uk.pdf

[29] Siirtokelpoinen tiedostomuoto on sellainen, jonka PAS-palvelu hyväksyy vastaanotossa ja jossa se muunnetaan tarvittaessa säilytyskelpoiseen tiedostomuotoon. Kaikki säilytyskelpoiset tiedostomuodot ovat myös siirtokelpoisia, mutta kaikki siirtokelpoiset eivät ole säilytyskelpoisia.

[30] http://ukdataservice.ac.uk/manage-data/format/recommended-formats

[31] PRONOM (https://www.nationalarchives.gov.uk/PRONOM/Default.aspx) on tiedostomuotokirjasto. Tiedosto-muotokirjasto on järjestelmä joka kuvaa tiedostomuotoja sekä sovelluksia, joilla ne ovat avattavissa ja/tai muunnettavissa tiedostomuodosta toiseen (migraatio). Se voi kuvata myös sovelluksien ominaisuuksia ja bugeja, jotka vaikuttavat näihin muunnoksiin laatua heikentävästi eli siten, että syntyvän tiedoston sisältö ja / tai ulkoasu ei vastaa alkuperäistä.

[32] http://www.kdk.fi/index.php/fi/pitkaaikaissailytys/maeaerittely-ja-dokumentit/5-suomi/pitkaeaikaissaeilytys/309-kayttoonottoprosessi

[33] JHove on työkalu, jolla tiedostoja voidaan tunnistaa ja osittain validoida: http://sourceforge.net/projects/jhove/

vaatimuksia, tiedon tallettajaa pyydetään korjaamaan viat. DANS tarjoaa myös tähän liittyvää koulutusta erilaisille käyttäjäryhmille[34].

NorStore vaatii, että käytetään sen määrittelemää metatietoskeemaa, kun aineisto lähetetään julkaistavaksi. Jos aineistoa ei julkaista kolmen kuukauden kuluessa siitä, kun se on siirretty vastaanottoon, aineisto poistetaan[35]. Jos palvelu huomaa aineiston olevan virheellistä tai puutteellista, tallettajalle annetaan kuukausi aikaa korjata se. Jos näin ei tapahdu, palvelu poistaa aineiston.

UKDAlla tiedostot tarkistetaan valmisteluvaiheen neuvotteluissa (*pre-ingest negotiations*) ja validoidaan vastaanotossa. Jos aineisto ei noudata määrityksiä, UKDA pyrkii korjaamaan aineiston itse; tarvittaessa aineisto palautetaan sen tallettajalle.

CINES linkittää julkaisut, tutkimusdatan ja menetelmät luomalla linkityksen säilytyspakettien välille. DANS toteuttaa linkityksen DataVerse-järjestelmässä, jossa tutkija voi linkittää tutkimukseen liittyviä komponentteja. Lisäksi NARCIS-järjestelmä sisältää tietoa julkaisuista, tutkimusdatasta ja menetelmistä.

NorStoren tavoitteena on tutkimusdatan DOI:n[36] yhdistäminen julkaisujen DOI:den kanssa käyttäen kansallista tieteellisten julkaisujen rekisteriä[37]. UKDA kerää tutkimusmenetelmät osana aineistojen metatietoja ja tarjoaa viitteet liittyviin julkaisuihin.

[34] http://datasupport.researchdata.nl/en
[35] Vastaanottoprosessi: https://archive.norstore.no/user-guide.pdf
[36] http://www.doi.org/
[37] http://www.cristin.no/english/

5 BITTITASON SÄILYTYS

Bittitason säilyttäminen -osiossa selvitettiin erityisesti bittitason säilyttämisen teknistä rakennetta. Tutkimuksen PAS-palvelussa bittitason säilyttäminen perustuu samaan ratkaisuun kuin KDK:n PAS-palvelussa, joten tutkimuksen PAS-palvelun kannalta toteutettuun ratkaisuun todennäköisesti ei ole tarpeellista tehdä muutoksia bittitason säilytyksen osalta. Tämän osion avulla kuitenkin pystymme osittain vertaamaan toteutettua PAS-ratkaisua valittujen organisaatioiden vastaaviin ratkaisuihin.

CINESin palvelulla on sekä DSA- että ISO16363-sertifikaatit[38]. Lisäksi CINESin auditoi joka kolmas vuosi ministeriöiden välinen instituutti (SIAF for Service interministériel des archives de France). DANSin palvelulla on vain DSA-sertifikaatti. NorStorella ei ole sertifikaatteja, mutta he pyrkivät saamaan DSA-sertifikaatin vuoden 2016 aikana. UKDAlla on DSA, ISO27000- ja ISO16363 -sertifikaatit (ISO16363 on itse auditoitu).

Taulukko 4: Palveluiden sertifikaatit

	DSA	NESTOR	ISO 16363	ISO 27000	Itse auditointi
CINES	X		X		X
DANS	X				
NorStore					
UKDA	X		(X)	X	
KDK-PAS				X	

Yhdelläkään kyselyyn osallistuneella palvelulla ei ole käytössään tallennustilakiintiötä tallettajille. CINESissä tallettajat toimittavat alkuvuodesta CINESille tiedon ennakoimastaan datan ja siirtopakettien määrästä. Jos määrä on suurempi kuin suunniteltu, tallettaja ottaa yhteyttä CINESiin varmistaakseen, että suurempi datamäärä on mahdollista siirtää säilytykseen.

DANS perustelee kiintiöiden puuttumista sillä, että kerralla talletettavat datamäärät ovat pieniä (<5Gb). Suuremmat datamäärät DANS:in palvelussa liittyvät yleensä projekteihin, joiden datamäärät tiedetään etukäteen datan hallintasuunnitelman perusteella ja joihin siten niihin osataan varautua. UKDA:lla on käytännön rajoituksia, mutta käytössä ei ole varsinaisia kiintiöitä.

CINES varmistaa datan eheyden vähintään vuosittain säilytyspakettien tarkistussummiin perustuen. DANS on ulkoistanut varsinaisen bittitason säilytyksen, joten DANSin tapauksessa eheyden varmistaminen perustuu palvelutasosopimukseen (SLA) sen ja säilytyksestä huolehtivan organisaation välillä. NorStoressa datan eheys varmistetaan iRODSin sisäisillä tarkistussummilla. UKDA ilmoittaa ISO 27000-sertifioitu. (27000-sarjan standardit käsittelevät tietoturvan eri aspekteja.)

CINESin tallennusarkkitehtuuri[39] luottaa tallennuksessa kiintolevyihin ja magneettinauhoihin. Sillä on kaikesta aineistosta kaksi kiintolevykopiota ja kaksi magneettinauhakopiota. Toinen magneettinauha-kopioista säilötään pimeään arkistoon. CINES toimii toistaiseksi yhden säilytyspisteen mallilla. Sillä on

[38] ISO16363 sertifikaatteja ei virallisesti vielä myönnetä, mutta CINES osallistui ISO16363 testi auditointiin APARSEN projektissa (ks. APARSEN deliverable D33.1)

[39] https://www.cines.fr/en/long-term-preservation/production-platform-2/technical-responses/hardware-solution-2/

suunnitelmissa ottaa käyttöön toinen säilytyspiste[40]. CINESin nykyinen tallennusratkaisu lähestyy elinkaarensa loppua, joten lähiaikoina on odotettavissa uutisia uudesta ratkaisusta.

DANS on ulkoistanut varsinaisen bittitason säilyttämisen, heillä ei ole palvelun sisäistä tallennusarkkitehtuuria. NorStoren tallennusarkkitehtuuri[41] luottaa säilytyksessä kiintolevyihin ja magneettinauhoihin – yksi kopio kumpaakin lajia. NorStore toimii yhden säilytyspisteen mallilla.

UKDA:n tallennusarkkitehtuurista[42] eivät valitut mediaratkaisut käy ilmi, mutta kopiota heillä on kaikesta aineistosta kolme kappaletta; säilytyspisteitä on kaksi. Sekä CINES että NorStore käyttää kiintolevyjä säilytyksessä olevien aineistojen jakeluun. Magneettinauhat toimivat rinnakkaisina säilytysmedioina, joilla ei palvella loppukäyttäjiä.

KDK:n PAS-palvelussa aineistosta säilytetään kolme kopioita eri mediatyypeillä ja lisäksi aineistosta säilytetään yhtä kopiota pimeässä arkistossa. Tämä strategia on verrannollinen kyselyyn osallistuneiden organisaatioiden kanssa. KDK:n PAS-palvelu on auditoitu ISO 27001-sertifikaatin osalta ja itseauditoitu niin kutsutun TRAC-tarkastuslistan (Trusted Repository Audit & Certification Checklist) avulla. Palvelun auditointisuunnitelma pitää sisällään DSA- ja ISO 16363 -sertifikaatit vuosien 2016–2018 aikana.

[40] https://www.cines.fr/en/long-term-preservation/production-platform-2/technical-responses/logical-architecture/
[41] https://www.sigma2.no/content/norstore-hardware-resources
[42] http://www.data-archive.ac.uk/media/3726/D6-HLH-ArchivalStorage-MultiCopyResilience_02_00.png

6 YMMÄRRETTÄVYYDEN SÄILYTTÄMINEN

Ymmärrettävyyden säilyttäminen -osiossa pyrimme selvittämään palveluiden lähestymistapaa semanttisen ja loogisen tason säilyttämiseen. Näillä säilyttämisen tasoilla varmistetaan, että säilytyksessä oleva aineisto ei ainoastaan ole teknisesti käytettävissä vaan että kohdeyhteisö[43] pystyy sen sisällön myös ymmärtämään vuosikymmenienkin kuluttua.

CINESillä kohdeyhteisön seuraamisesta vastaa aineiston tallettaja. Tämä tapahtuu pitämällä yllä PPDI (Project Preservation Description Information)[44] -dokumenttia. CINESillä ei ole prosessia, jolla reagoidaan kohdeyhteisössä tapahtuviin muutoksiin.

DANSin palvelussa kohdeyhteisön seuraamisesta vastaa säilytyspalvelu. Tämä tapahtuu osallistumalla projekteihin ja yksilöimällä kohdeyhteisöt datan hallintasuunnitelmissa. DANS pyrkii muutoksien tapahtuessa mukauttamaan prosessejaan ja järjestelmiään.

NorStorella kohdeyhteisön seuraamisesta vastaa ”Advanced user support”. Toiminnon on tarkoitus toteutua vuorovaikutuksessa yhteisöjen kanssa, mutta toistaiseksi tätä toimintoa ei ole täysin vakiinnutettu. Muutoksien tapahtuessa, NorStore ottaa yhteyttä muutoksien kohteena olevien aineistojen hallinnasta vastaavaan tahoon.

UKDA:ssa kohdeyhteisön seuraamisesta vastaa säilytyspalvelu. UKDA:lla on määritelty muutoksiin reagoinnin prosesseja (*Business misen prosessejaprocess and metadata change management, technical change management*), mutta niitä ei ole vastauksissa tarkemmin kuvattu.

Säilytyssuunnitelman alkuperäisen tekijät ja ylläpitäjät on esitetty taulukossa 5. CINESillä säilytyssuunnitelmat perustuvat PPDI-dokumenttiin/skeemaan. DANSilla säilytyssuunnitelman oletetaan olevan osa datan hallintasuunnitelmaa. NorStorella alkuperäisestä säilytyssuunnitelmasta vastaa aineiston tallettaja eikä heillä toistaiseksi ole suunnitelmia säilytyssuunnitelmien ylläpitämiseksi. UKDA:lla säilytyssuunnitelmista vastaa palvelu sekä aineistoa talletettaessa että säilytyksen aikana.

Taulukko 5: Vastuut säilytyssuunnitelmista

	Alkuperäisen säilytyssuunnitelman tekee	Säilytyssuunnitelmaa ylläpitää
CINES	Tallettaja	Palvelu
DANS	Palvelu	Palvelu
NorStore	Tallettaja	Ei päätetty
UKDA	Palvelu	Palvelu

[43] Kohdeyhteisö on tietty ryhmä potentiaalisia käyttäjiä, joiden pitäisi pystyä ymmärtämään tiettyä tietokokonaisuutta.
[44] https://www.cines.fr/pac/ppdi.xsd

7 KÄYTTÖ

Käyttöä koskevassa osiossa selvitettiin,

- kuka kontrolloi säilytyksessä olevaa aineistoa,
- kenellä siihen on pääsy,
- kuka pääsystä päättää ja
- kuka saa poistaa säilytyksessä olevaa aineistoa.

Aineiston säilyttämisen lähtökohtana on aineiston käytettävyys tulevaisuudessa. Käytettävyydellä tarkoitetaan tässä yhteydessä sitä, että aineisto on mahdollisimman helposti ja mutkattomasti, ilman teknisiä tai hallinnollisia välivaiheita, mahdollisimman laajan asiakaskunnan tarkasteltavissa ja uudelleen käytettävissä. Käytettävyyden varmistamiseksi tulee säilytyksessä olevan aineiston kontrollista, käyttöoikeuksista ja lisensseistä sopia selvästi aineiston vastaanottovaiheessa.

Kartoituksessa tunnistettiin kahdenlaisia palveluita. Palvelu voi joko (1) keskittyä ainoastaan aineiston säilyttämiseen tai (2) säilyttämisen lisäksi myös aineiston käytettävyyden varmistamiseen. Käytettävyyden varmistaminen alkuperäistä tallettajaa laajemmalle käyttäjäjoukolle edellyttää huomattavasti tarkempaa kontrollia ja käyttöoikeuksien määrittelyä vastaanottovaiheessa. Mikäli aineistoa säilytetään ainoastaan alkuperäisen tallettajan käyttöön, lisenssit tai käyttöoikeudet eivät ole säilyttävän palvelun huoli.

Taulukko 6: Säilytyksessä olevan aineiston saatavuuden kontrollointi

	Tallettaja kontrolloi	Palvelu kontrolloi
CINES	X	
DANS		X
NorStore		X
UKDA		X
KDK-PAS	X	

CINESillä aineiston saatavuutta kontrolloivat aineiston tallettajaorganisaatiot, jotka ovat sopineet aineiston tuottajan kanssa sen käyttöoikeuksista. CINES valvoo lisenssien noudattamista siten, että aineistoa tallettava organisaatio sopimuksella hyväksyy lisenssien noudattamisen aineiston tuottajan kanssa perustuen kansalliseen lainsäädäntöön. Heidän toimintamallissaan tallettava organisaatio on siis keskeisessä roolissa aina aineiston valinnasta alkaen ja kontrolloi myös CINESissä olevaa aineistoa. Aineiston poistamisesta säilytyksestä CINESillä päättää pääasiallisesti tallettaja; joissakin tapauksissa myös palvelu voi tehdä päätöksen.

DANSilla saatavuutta kontrolloi palvelu, mutta tallettaja voi vaatia tietoa aineiston käyttötarkoituksesta ennen kuin käyttölupa myönnetään. Palvelu vastaa käyttötarkoituksen tiedottamisesta tallettajalle. DANSilla lisenssit ovat osa aineiston dokumentaatiota ja hankintaprosessia. DANS pyrkii noudattamaan periaatetta ”*avointa jos mahdollista, suljettua tarvittaessa*”.

NorStorella puolestaan teknistä saatavuutta kontrolloi palvelu. Koska palvelu hyväksyy ainoastaan avointa dataa, käyttörajoituksia ei ole. NorStorella ei ole toistaiseksi prosessia lisenssien noudattamisen valvomiseksi.

UKDA:lla aineiston saatavuus riippuu siitä, mitä tallettajan ja palvelun kanssa on tapauskohtaisesti sovittu. UKDA:lla aineistojen käyttöä tarkkaillaan tiukasti. Sillä on erilaisia menetelmiä lisenssirikkomusten havaitsemiseen, vaikkakin keinot käytännössä ovat rajalliset erityisesti avoimen tutkimusdatan osalta.

KDK:n PAS-palvelun tapauksessa palvelun rooli on rajattu ainoastaan säilyttämiseen ja aineiston jakelu on mahdollista vain ainoastaan alkuperäiselle tallettajalle (hyödyntävälle organisaatiolle). Tämä mahdollistaa sen, että palvelun ei tarvitse ottaa kantaa säilytettävien aineistojen käyttöoikeuksiin. Tutkimusaineistojen tapauksessa jakelu on tunnistettu tärkeäksi käyttötapaukseksi. Tutkimusaineistojen jakelun siirtäminen tutkijoilta keskitetyn palvelun tehtäväksi siirtää asiaan liittyvän teknisen kuorman pois tutkijoilta ja tutkimuslaitoksilta.

Taulukko 7: Säilytyksessä olevan aineiston poistaminen

	Tallettaja poistaa	Palvelu poistaa
CINES	X	X
DANS	X	
NorStore		
UKDA	X	
KDK-PAS	X	

LIITE A: SÄILYTYS- JA SIIRTOKELPOISET TIEDOSTOMUODOT

Seuraavassa taulukossa on lueteltu kyselyyn osallistuneiden palveluiden ja KDK:n PAS-palvelun[45] hyväksymät säilytys- ja siirtokelpoiset tiedostomuodot (A=säilytyskelpoinen; B=siirtokelpoinen). Taulukko on vain suuntaa antava; esimerkiksi tiedostomuotojen versioita ei ole huomioitu. Lisäksi esimerkiksi UKDA mainitsee osan tiedostomuodoista vain esimerkkinä, joten he hyväksynevät myös muita tiedostomuotoja. Lisäksi on huomioitava, että KDK:n PAS-palvelussa on keskitytty ainoastaan kulttuuriperintöaineistoihin ja siten tutkimusaineistojen tarvitsemat tiedostomuodot on sivuutettu täysin.

		KDK	CINES	DANS	UKDA
Teksti	EPUB	A			
	XHTML	A			B
	XML	A	A	B	A
	HTML	A	A	B	A
	ODT	A	A	B	A
	PDF/A	A	A	A	A
	TXT	A	A	A	A
	PDF	B	A	B	A
	DOC(X)	B		B	B
	RTF			B	A
	NUD*IST				B
	NVivo				B
	ATLAS.ti				B
Ääni	AIFF	A	A		B
	BWF	A			
	FLAC	A	A		A
	MPEG-4 AAC	A	A		
	WAV	A	A	A	B
	OGG		A		
	AIFF-C	B			
	MP3	B		B	B
Elävä kuva	JPEG2000	A			A
	MKV		A		
	MPEG-2	B		A	
	MPEG-4	A	A	A	A
	AVI			A	
	MOV			A	
	DV	B			
	WMV	B			
Kuva	DNG	A			
	JPEG	A	A	A	B
	JPEG2000	A	A	B	
	PNG	A	A	A	
	TIFF	A	A	A	A
	GIF	B	A		
	EPS	B		B	
	SVG		A	A	

[45] http://www.kdk.fi/index.php/fi/pitkaaikaissailytys/maeaerittely-ja-dokumentit/5-suomi/pitkaeaikaissaeilytys/141-kdkn-saeilytys-ja-siirtokelpoiset-tiedostomuodot

		KDK	CINES	DANS	UKDA
	RAW				B
	PSD				B
	AI			B	
Taulukkolaskenta	ODS	A		A	B
	CSV			A	
	XLS(X)	B		A	B
Tietokannat	ANSI SQL			A	
	CSV			A	B
	dBase			B	B
	Access			B	B
Tilastotiede	R			A	
	SPSS Portable			A	A
	SAS Transport			A	A
	STATA			A	A
CAD	AutoCAD DXF			A	
	AutoCAD others			B	
GIS	GML			A	
	MapInfo interchange			A	B
	GeoTIFF		A	A	A
	ESRI Shapefiles			B	A
	MapInfo			B	
	KML			B	B
	TIFF World File			B	
	CAD data (dwg)				A
	ESRI Geodatabase				B
	Adobe AI				B
	CAD data (dxf. svg)				B
	ASCII GRID			A	
	ESRI GRID			B	

LIITE B: KYSELYLOMAKE

OPEN SCIENCE AND RESEARCH

Survey on Research Data Preservation Practices

Open Science and Research Initiative of Finland is conducting a survey of selected digital preservation repositories for research data to find out current best practices and lessons learned. We are mainly interest in your policies and processes for preserving research data than the technical implementation, although we would like to know insights of your technical implementation as far as possible.

Please take some time to complete this questionnaire and share your views on research data digital preservation repositories with us. The questionnaire is split into 6 sections: (1) Background, (2) Acquisition, (3) Ingest, (4) Storage, (5) Content preservation, and (6) Access.

Although almost all questions are optional, we would like to have your answer to all of them. Your link to this questionnaire is personal (but you can share it with your colleagues), and therefore you do not have to complete the whole questionnaire at once, but you can continue at any time later until you submit the questionnaire. Further, the most of the questions have an optional free form text field for providing further details related to the question. Please use this field to give as detailed information as possible. Give URL's to relevant documentation where appropriate.

Should you have any questions, do not hesitate to contact Heikki Helin.

Please complete the questionnaire before **June 30th, 2015**. Note that this questionnaire is sent only to a few repositories and thus your answers are crucial. We thank you for your participation.

Best regards,
The Digital Preservation team of the Open Science and Research Initiative of Finland.

Background

In this section, we would like to have a general overview of your repository and its purpose.

1. **Name of the repository***

   _____

2. **Repository URL**

   _____

3. **Personal details and contact information**

   Name (of person completing questionnaire)* _____
   Email* _____
   Job title _____

4. Type of the organization?

- ☐ Academic/research library
- ☐ Data centre
- ☐ Government department
- ☐ Institutional archive
- ☐ Institutional library
- ☐ Museum
- ☐ National archive
- ☐ National library
- ☐ Regional archive
- ☐ Regional library
- ☐ Research unit (e.g. university department)
- ☐ Other, please specify _____

5a. What kind of data your repository contains?

- ☐ Publications
- ☐ Processed data
- ☐ Raw research data
- ☐ Other, please specify _____

5b. Please, provide further details for the question above

6. From which research disciplines the data is from?

Especially if there are any restrictions, please describe those in detail.

7a. How much data your repository currently contains (in terabytes excluding copies)?

_____

7b. How much data you anticipate to have in by the end of 2020 (in terabytes excluding copies)?

_____

8. Describe the governance structure of your repository as detailed as possible. What kind of roles you have and what are their responsibilities?

9a. Do you have any cooperation with other repositories?

- ○ Yes, cooperation is in place.
- ○ No, but we are planning to do that.
- ○ No

9b. Please, provide further details for the question above

Acquisition

In this section, we would like to know how the data to be stored in your repository is selected, for how long the data is preserved, who owns the data after it is stored in the repository, and grounds for the sustainability of the repository.

10a. Who stores data in your repository?

10b. Regarding the question above, who selects the depositors which are able to store data in your repository?

11a. Who selects the data to be preserved?

11b. Regarding the question above, on what ground the selection is done?

12. Who owns the preserved data?

- ☐ Depositor
- ☐ Funder
- ☐ Repositor
- ☐ Not applicable
- ☐ Other, please specify _____

13a. For how long should the data be preserved (by default)?

Please select multiple options if planned preservation time depends on the data.

- ☐ Max. 5 years
- ☐ 5 – 20 years
- ☐ 20 – 50 years
- ☐ 50 – 100 years
- ☐ Over 100 years
- ☐ Not decided
- ☐ Not known
- ☐ Not applicable

13b. Please, provide further details for the question above

14a. Is there dedicated funding for the actions/activities in the repository?

- ☐ Annual public funding
- ☐ Funded by commercial funders
- ☐ Funded by depositors
- ☐ Funded by non-profit funders
- ☐ Funded by public funders
- ☐ Funded by sales activities
- ☐ Lump sum public funding
- ☐ Short-term project-based funding
- ☐ Not applicable
- ☐ Other, please specify _____

14b. Please, provide further details for the question above

15. Are there any incentives for depositors to store their data?

- ○ No
- ○ Yes, please specify _____

16a. Are licenses required for the data to be preserved by the repository?

- ☐ Copyright protected, no licenses
- ☐ International licenses
- ☐ National licenses
- ☐ Public domain
- ☐ Open licenses (e.g. Creative Commons)
- ☐ Uniform collection of licenses of various types
- ☐ Not applicable
- ☐ Other, please specify _____

16b. Are licenses recommended for the data to be preserved by the repository?

- ☐ Copyright protected, no licenses
- ☐ International licenses
- ☐ National licenses
- ☐ Public domain
- ☐ Open licenses (e.g. Creative Commons)
- ☐ Uniform collection of licenses of various types
- ☐ Not applicable
- ☐ Other, please specify _____

17. Describe your process for making sure that the license policy is complied with?

Ingest

In this section, we would like to find out insights what kind of technical requirements your repository have for the data and its metadata, and how these are guaranteed before the data is actually taken into preservation.

Requirements for metadata

What kind of metadata you require with the data.

18. Your requirement for metadata format for packaging the data (e.g. METS) is... (choose applicable option)

If any requirements apply, please provide details of acceptable formats and other requirements. If you don't have any requirements, please provide insights why not.

○ Mandatory ○ Recommended ○ Optional ○ Not applicable

19. Your requirement descriptive metadata is... (choose applicable option)

If any requirements apply, please provide details of acceptable formats and other requirements. If you don't have any requirements, please provide insights why not.

○ Mandatory ○ Recommended ○ Optional ○ Not applicable

20. Your requirement provenance metadata is... (choose applicable option)

If any requirements apply, please provide details of acceptable formats and other requirements. If you don't have any requirements, please provide insights why not.

○ Mandatory ○ Recommended ○ Optional ○ Not applicable

21. Your requirement technical metadata is... (choose applicable option)

If any requirements apply, please provide details of acceptable formats and other requirements. If you don't have any requirements, please provide insights why not.

○ Mandatory ○ Recommended ○ Optional ○ Not applicable

22. Your requirement rights metadata is... (choose applicable option)

If any requirements apply, please provide details of acceptable formats and other requirements. If you don't have any requirements, please provide insights why not.

○ Mandatory ○ Recommended ○ Optional ○ Not applicable

23. Your requirement structural metadata is... (choose applicable option)

If any requirements apply, please provide details of acceptable formats and other requirements. If you don't have any requirements, please provide insights why not.

○ Mandatory ○ Recommended ○ Optional ○ Not applicable

24. Are there any other metadata requirements you would like to share with us?

Requirements for file formats

Some repositories have a selected set of file formats as others allow data to be in any format.

25. Do you have requirements for content data file formats?

Please provide URLs to appropriate documentation if possible.

- ○ No, please specify reasons why not:
- ○ Yes, please specify which formats are allowed:

26. If you have requirements for file formats, what kind of process you have for updating the requirements?

Quality assurance in ingest

If you have requirements for the metadata and/or file formats (regarding questions above).

27. How do you make sure that metadata/format requirements given above are taken into account when ingesting data?

28. What is the procedure when the requirements are not satisfied?

Managing the research data as a whole

Research data typically needs various supplement content information with it, in order to create a whole which can be understood/interpret and especially that it can be utilized later (after several, or tens of, years later).

29. Do you have any requirements/plans how to link research data, publications and research methods together?

Storage

Repository admin & audits

30. How the technical administration of the repository is organized?

Please, specify roles and responsibilities as detailed as possible.

31. Is your repository certified?

- ☐ Yes, we have Data Seal of Approval (DSA)
- ☐ Yes, we have a Nestor certificate
- ☐ Yes, we have ISO 16363
- ☐ Yes, we have certificate(s) from ISO 27000 series. Please specify:
- ☐ No, but we have done self-auditing. Please specify:
- ☐ No, but we are planning to do that. Please specify:
- ☐ No, not applicable. Please specify:

Storage

32a. Are there any restrictions for how much data a depositor can store (e.g, quota)?

- ○ Yes
- ○ No

32b. Please, provide further details for the question above

33. How the repository makes sure that the data remains intact?

For example, security, access control, logs, ...

34. How often the data is refreshed, i.e., copied to a new media?

35. Please, provide details of your storage architecture

Content preservation

In this section, we would like to find out insights how the semantic and the logical preservation are taken care of in your repository.

Preservation watch

Following what is going on in designated communities is fundamental for digital preservation for a long-term.

36. Who is responsible, in your repository, for following the designated communities?

- ○ Repository
- ○ Depositor
- ○ Funder
- ○ Not applicable
- ○ Other, please specify: [ ]

37. How the designated communities are followed?

[ ]

38. What kind of process you have, if something changes in the designated community and/or the environment of your repository?

[ ]

Preservation planning

Preservation planning is needed in order to make sure that any material in a repository remains usable over time no matter how the environment (hardware, software, designated communities, etc) will change.

39. Who is mainly responsible for the initial preservation plan when the data is ingested?

- ○ Repository
- ○ Depositor
- ○ Funder
- ○ Not applicable
- ○ Other, please specify: [ ]

40. Who is mainly responsible for updating preservation plans when the data has been stored for some time (5-10 years)?

- ○ Repository
- ○ Depositor
- ○ Funder
- ○ Not applicable
- ○ Other, please specify: [ ]

41. What kind of requirements you have for preservation plans?

This question covers both technical and non-technical requirements.

42a. Are you planning any file format migrations for the content information in your repository?

- ○ Yes, doing that already
- ○ Yes, later
- ○ No
- ○ Not applicable

42b. Please, provide further details for the question above

Disposal

Sometimes it is needed to dispose some data from the repository although originally considered to be preserved for a long time.

43. In your repository, who makes the decision to dispose any data?

- ○ Depositor
- ○ Funder
- ○ Repository
- ○ Not applicable
- ○ Other, please specify:

44. Have you had any cases in which data was disposed unexpectedly (e.g, human errors, technical errors)?

- ○ Yes, please specify:
- ○ No

Access

45a. Who controls the access to data?

- ○ Funder
- ○ Not applicable
- ○ Other, please specify:

45b. Please, provide further details for the question above

For example, what is your process to handle cases when somebody wants access to data that has some usage restrictions.

46. Please, provide any other information regarding your repository (e.g., URLs to documentation).