

Kansallinen digitaalinen kirjasto: Digitaalinen pitkäaikaissäilytys

Arto Teräs <arto.teras@csc.fi>
Ajankohtaisseminaari digitoinnista
Kansallismuseo 11.2.2009



Esityksen sisältö

- **Mitä pitkäaikaissäilytys tarkoittaa?**
- **Pitkäaikaissäilytys osana opetusministeriön Kansallinen digitaalinen kirjasto (KDK) -hanketta**
- **Pitkäaikaissäilytysjärjestelmän (PAS-järjestelmän) alustavan toiminnallisen vaatimusmäärittelyn pääkohdat**
 - Kokonaisuus, toimijat ja roolit
 - Arkistojen, kirjastojen ja museoiden aineisto ja asiakkaat
 - PAS-järjestelmän suhde olemassaoleviin toimintoihin
 - Säilytettävän informaation luovutus PAS-järjestelmään
 - Tekninen infrastruktuuri
 - Turvallisuus ja riskit
 - Standardit, lait ja normit
- **Lisätyötä vaativat kokonaisuudet**



Mitä pitkäaikaissäilytys (PAS) tarkoittaa?

- **Digitaalisen informaation luotettavaa säilyttämistä useiden kymmenien tai satojen vuosien ajan**
 - Laitteet ja ohjelmistot, jopa tiedostomuodot vanhenevat, mutta informaation täytyy säilyä
 - Luotettava säilyttäminen edellyttää sisällön eheyden aktiivista valvontaa ja monenlaisiin riskeihin varautumista
- **Informaation, ei pelkästään bittien säilyttämistä**
 - Metatiedot tärkeässä asemassa
- **Informaation säilyttämistä jopa pitempään kuin joidenkin organisaatioiden elinaika**
 - Tarvitaan yhteistyötä, standardeja ja käytännöistä sopimista



Kansainvälisen katsauksen havaintoja

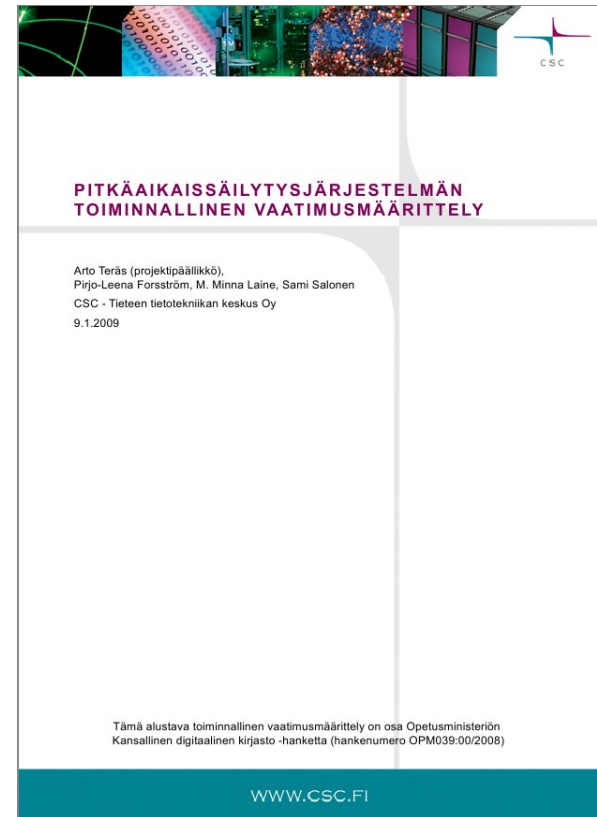
Pitkäaikaissäilytyksen projektit hyötyvät seuraavista asioista:

- **Selkeä organisoituminen ja vastuutus**
- **Standardien noudattaminen**
- **Kansainvälinen yhteistyö**
- **Teknologiainfrastruktuurin syvällinen tuntemus**
- **Exit-strategiat, eli valmiudet muuttaa tehtyjä ratkaisuja ja siirtyä uuteen toimintamalliin tilanteen muuttuessa**
 - Esimerkiksi jos tietojärjestelmän myynyt taho lopettaa toimintansa



Pitkäaikaissäilytys KDK-hankkeessa

- **Tavoite: Pitkäaikaissäilyttämisen ratkaisu arkistojen, kirjastojen ja museoiden keskeisille sähköisille tietovarannoille**
 - Organisaatioiden käyttöön tarjotaan toimintaa tukeva pitkäaikaissäilyttämisen ratkaisu
 - Organisaatioiden tietojärjestelmät liitetään PAS-järjestelmään avoimilla rajapinnoilla
 - Toimijat vastaavat tulevaisuudessakin omien digitaalisten aineistojensa tuottamisesta ja hallinnasta
- **Toteutusaika 1.6.2008 –30.5.2010, jolloin valmiina PAS-järjestelmän toiminnallinen ja tekninen vaatimusmäärittely**
 - PAS-järjestelmän toteuttaminen erikseen
- **Ensimmäisenä työkokonaisuutena on laadittu alustava toiminnallinen vaatimusmäärittely**



Suunnittelun lähtökohdat

- **Jokaisella toimijalla on omat aineistot, niiden erityispiirteet ja niihin liittyvät metatiedot**
- **Jokaisella on omat tiedon tuottamisen tietojärjestelmät**
- **Pitkäaikaissäilytysratkaisun on kuitenkin tarkoitus palvella kaikkia — tarvitaan yhteisesti sovittuja toimintatapoja**
 - Tiedostomuotojen ja metatietojen yhtenäistäminen
 - Yhteiset pelisäännöt, missä vaiheessa ja miten metatiedot lisätään aineistoon
 - Sovitaan missä ja miten mahdollisesti tarvittava aineiston muokkaus pitkäaikaissäilytystä varten tehdään, jotta päästään mahdollisimman tuottavaan ratkaisuun
- **Edetään yhteisistä tavoitteista kohti omia erityisratkaisuja**
- **Päätavoitteena digitaalisen informaation turvallinen säilytys**
 - Samalla mahdollisuus tehostaa toimintoja ja prosesseja, kehittää yhteistyötä ja yhdenmukaistaa tietojärjestelmiä, saaden sitä kautta kustannussäästöjä ja parempia palveluja kansalaisille

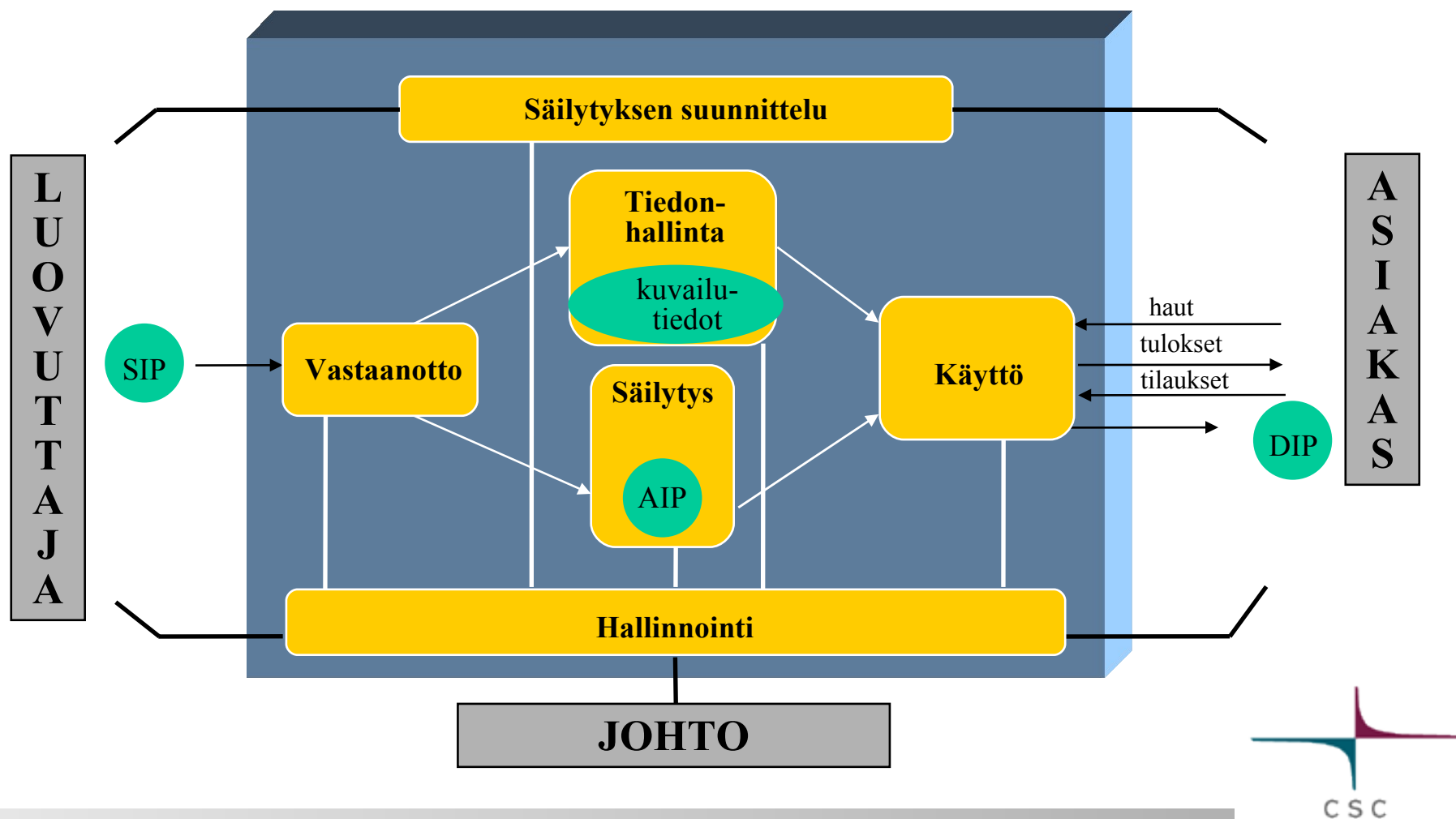


Toiminnallisen vaatimusmäärittelyn lähtökohdat

- **OAIS-viitemalli (Reference Model for an Open Archival Information System, tammikuu 2002)**
 - Myös OAIS:n suomenkielisen käännöksen luonnos huomioitu
- **TRAC-tarkistuslista (Trustworthy Repositories Audit & Certification: Criteria and Checklist, v. 1.0, helmikuu 2007)**
- **Kansalliskirjaston vaatimusmäärittely pitkäaikaissäilytysjärjestelmälle, heinäkuu 2007**
- **Taustatietoina konferensseissa esitettyä materiaalia, artikkeleita sekä muita standardeja ja määrityksiä, mm.**
 - MoReq2 specification v. 1.04 - Update and extension of the Model Requirements for the management of electronic records, 2008
 - Kansallisarkiston vastaanotto- ja palvelujärjestelmän (VAPA) vaatimusmäärittely, marraskuu 2006
 - SÄHKE2 Asiakirjallisen tiedon käsittely, metatietomalli ja toiminnalliset vaatimukset, luonnos, lokakuu 2008



Pitkäaikaissäilytyksen kokonaisuus OAIS:n toiminnallisen mallin pohjalta

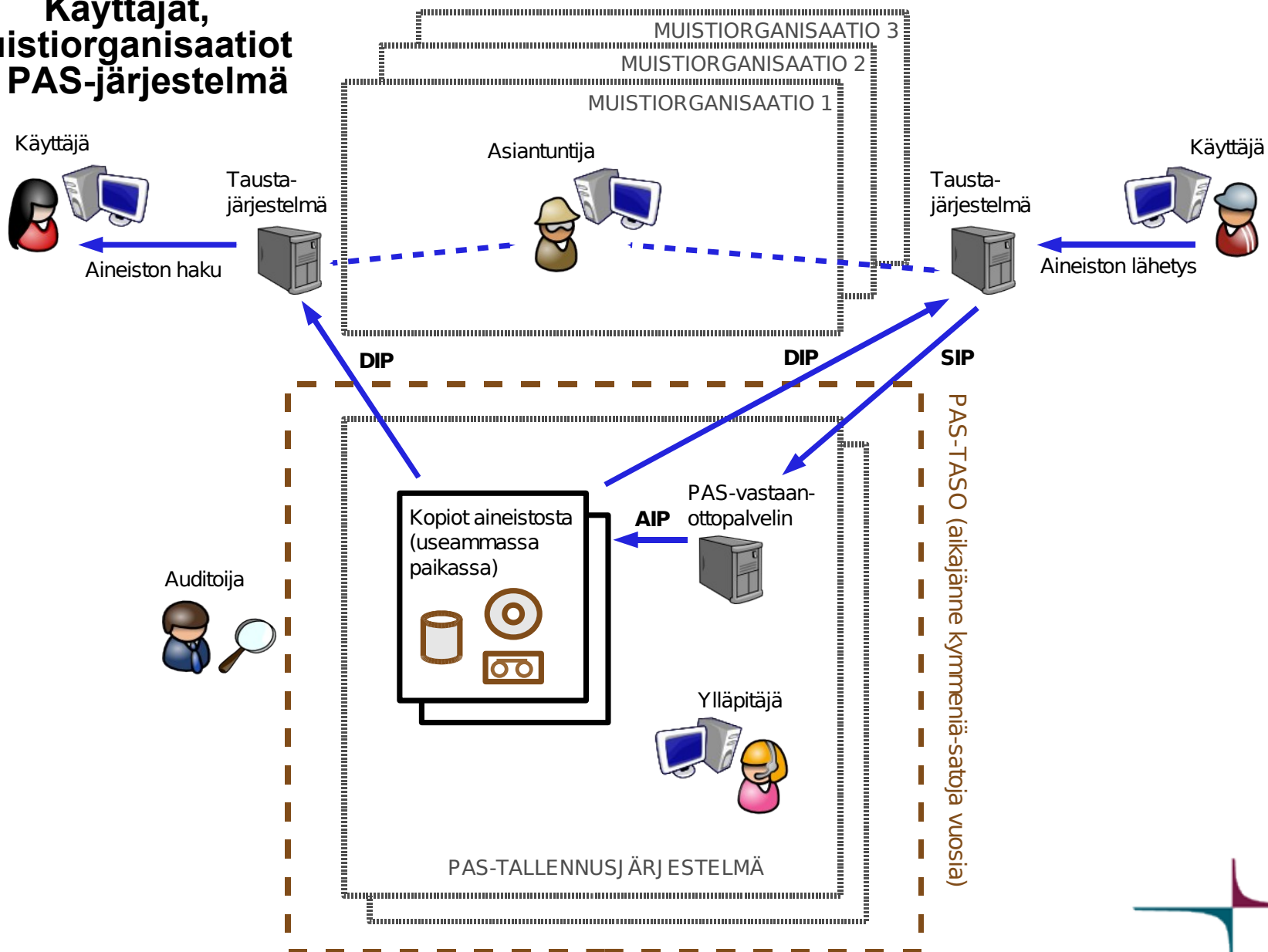


OAIS-mallin mukaiset tietopaketit

- **Luovutuspaketti** (Submission Information Package, **SIP**) koostuu luovutettavasta aineistosta ja siihen liittyvistä metatiedoista
- **Säilytyspaketti** (Archival Information Package, **AIP**) on PAS-järjestelmässä säilytettävä tietopaketti, joka muodostetaan yhden tai useamman luovutuspaketin pohjalta. Se koostuu sisällöstä ja siihen liittyvästä säilytysinformaatiosta
- **Jakelupaketti** (Dissemination Information Package, **DIP**) on yhdestä tai useammasta säilytyspaketista koostettu tietopaketti, jonka asiakas (tai taustajärjestelmä) saa tehtyään aineistopyynnön PAS-järjestelmälle
- **Tietopaketti** (Information Package) on yleistermi edellämainituista



Käyttäjät, muistiorganisaatiot ja PAS-järjestelmä



Toimijat

- **PAS-organisaatio** tarkoittaa koalitiota, joka ohjaa, vastaa ja osallistuu KDK-hankkeen pitkäaikaissäilytykseen
- **PAS-järjestelmää ylläpitävä taho** on tämän organisaation valitsema toimija(t), jo(t)ka vastaa(vat) operatiivisista palveluista
- **PAS-järjestelmä** on kokonaisuus jolla pitkäaikaissäilytys toteutetaan
- **Luovuttaja** on kirjasto, arkisto, museo tai muu taho, joka on laatinut sopimuksen PAS-organisaation kanssa luovuttamisesta ja on näin auktorisoitu luovuttaja
- **Kohdeyleisö** on PAS-organisaation määrittämä asiakas, jolle PAS-järjestelmän tulee kyetä tarjoamaan säilytyksessä olevaa digitaalista informaatiota



PAS-organisaation vastuulla on

- Laatia luovuttajien kanssa sopimukset säilytettävistä informaatioasisällöistä
- Hankkia pitkäaikaissäilytyksen edellyttämät oikeudet luovutettuun informaatioon
- Laatia riskianalyysi ja valita toimenpiteet uhkien minimoimiseksi
- Varmistaa luovuttajien riittävä koulutus ja osaaminen
- Valita luotettava PAS-järjestelmää ylläpitävä taho
- Määritellä joko itse tai yhteistyökumppaneiden kanssa PAS-järjestelmän kohdeyleisö
- Varmistaa että säilytetty informaatio on kohdeyleisön itsenäisesti ymmärrettävissä



PAS-järjestelmän ja sitä ylläpitävän tahon vastuulla on

- **Tarjota luovuttajalle riittävät työkalut luovutuspakettien tuottamiseen**
- **Ottaa vastaan luovutuspaketit, muodostaa niistä säilytyspaketit ja siirtää ne PAS-järjestelmään**
- **Noudattaa dokumentoituja toimintaperiaatteita ja -menetelmiä jotka varmistavat informaation säilymisen kaikkia tunnettuja uhkia vastaan**
 - Informaatiota voidaan jakaa joko autentikoituina kopioina tai alkuperäiseen dokumenttiin palautettavissa olevina versioina
- **Tarjota säilytetty informaatio kohdeyleisön käyttöön**
 - Tästä voi vastata myös luovuttaja, jolloin PAS-järjestelmä tarjoaa säilytetyn informaation takaisin luovuttajan tuotantojärjestelmiin



Luovuttajan (kirjaston, arkiston tai museon) vastuulla on

- **Tuottaa ja koota sopimuksen mukaisesti luovutuspaketti sisällöltään sopivaksi**
- **Varmistaa luovutuspaketin eheys, kattavuus ja oikeellisuus**
- **Siirtää luovutuspaketti sovitulla tavalla PAS-järjestelmän vastaanottoon**
- **Ylläpitää luovutuspakettien informaation sisältöä ja luovutuspaketteihin liittyvää kokoelmatietoa niin omaan kuin asiakasliittymän käyttöön**
- **Tuotantojärjestelmien liittäminen PAS-järjestelmään**



Kohdeyleisön vastuulla on

- **Asettaa minimivaatimukset metatiedolle, jotta PAS-järjestelmässä säilytettävää digitaalista informaatiota voidaan etsiä ja tunnistaa**
 - Yhteistyössä PAS-organisaation ja luovuttajien kanssa



Säilytettävä informaatio

<u>Muistiorganisaatio</u>	<u>Säilytettävä informaatio</u>	<u>Tiedostomuoto</u>
Arkistolaitos, Digitaaliarkisto	Maakuntien ja kirkon vanhoja asiakirjoja	tiff
Kansalliskirjasto	Suomalaisen internetin sisältöä; vanhoja sanomalehtiä ja kausijulkaisuja, monografioita, suomalaisen kirjallisuuden, nuottien ja äänitteiden luettelo, ym.	jpg, pdf, mahdollisesti useita
Kansallinen audiovisuaalinen arkisto (KAVA)	Teksti, ääni, kuva, elävä kuva, multimedia	useita
Yhteiskuntatieteellinen tietoarkisto	"Raaka" tutkimusaineisto (datamatriisit, teksti, puhe, kuva)	SPSS Portable
Elinkeinoelämän keskusarkisto (ELKA)	Yritysten, elinkeinoelämän järjestöjen ja elinkeinoelämässä vaikuttaneiden henkilöiden pysyvästi säilytettäviä asiakirjoja	useita (tiff, jpg, wav, mpeg2,...)
Yksityiset keskusarkistot	Yksityisiä asiakirjoja, valokuvia, perinneainestoja, ääni- ja kuvatallenteita, karttoja, piirustuksia.	useita (tiff, jpg, wav, mpeg2 jne.)
Tilastokeskus	Suomen viralliset tilastot, julkaisut, arkistot; alkuperäistilastot pääosin luottamuksellisia; Arkistolaitos säilöö	xml -> pdf, html + vanhat järjestelmät
Luonnontieteellinen keskusmuseo	Eläin- ja kasvinäytteet, luonnontieteellinen havaintoaineisto, kulttuuriaineisto (matkamuistiinpanot ym.)	mm. xml (standardi)
Museovirasto	Kulttuurihistoriallisten museoiden ja erikoismuseoiden kokoelmat	mahdollisesti useita
Valtion taidemuseo / kuvataiteen keskusarkisto	Kirjasto, taidehistorialliset asiakirjat-arkisto, kuva-arkisto, äänitearkisto	mahdollisesti useita, mm. wav
Valokuvataiteen museo	Digitaalisia valokuvia, tekstimuotoisia sisältöjä tietokannoissa	mahdollisesti useita

Arkistojen, kirjastojen ja museoiden asiakkaat

<u>Muistiorganisaatio</u>	<u>Asiakkaat</u>
Arkistolaitos, Digitaaliarkisto	Sukututkijat, historian tutkijat, yksityishenkilöt, viranomaiskäyttäjät, selvitystyöryhmät, yrityskäyttäjät, esim. asianajajatoimistot
Kansalliskirjasto	Humanististen ja yhteiskuntatieteiden tutkijat ja opiskelijat, kirjastot, korkeakoulut
Kansallinen audiovisuaalinen arkisto (KAVA)	Tallettajat (lakisääteiset ja vapaaehtoiset), elokuvan- ja ohjelmantekijät, tutkijat, toimittajat, kirjailijat, kansalaiset yleensä
Yhteiskuntatieteellinen tietoarkisto	STAKES, KTL, Tilastokeskus, KTK; akatemia erit. ihmis- ja sosiaalitieteet; kansainväliset tutkimusohjelmat
Elinkeinoelämän keskusarkisto (ELKA)	Arkistojen luovuttajat, historian tutkijat, museot, eri alojen harrastajat, sukututkijat, dokumenttien tekijät, näyttelyiden suunnittelijat
Yksityiset keskusarkistot	Arkistojen luovuttajat, kulttuuri-, historian- ja perinnetutkijat, museot, eri alojen harrastajat, kotiseutu- ja sukututkijat, dokumenttien tekijät, näyttelyiden suunnittelijat.
Tilastokeskus	Laajasti eri yhteiskuntaryhmät ja yksityiset henkilöt Suomesta ja ulkomailta
Luonnontieteellinen keskusmuseo	Kotimaiset ja ulkomaiset tutkijat
Museovirasto	Museoammattilaiset, tutkijat, perusopetus, viranomaiset, yritykset sekä yksittäiset kansalaiset
Valtion taidemuseo / kuvataiteen keskusarkisto	Museot, opiskelijat, kansalaiset, Valtion taidemuseon oma henkilökunta, koti- ja ulkomaiset kuvankäyttäjät
Valokuvataiteen museo	Museon sisäinen kokoelmanhallinta, museon tutkija-asiakkaat; suuri yleisö, kuvapalvelupyynnöt julkaisukäyttöön

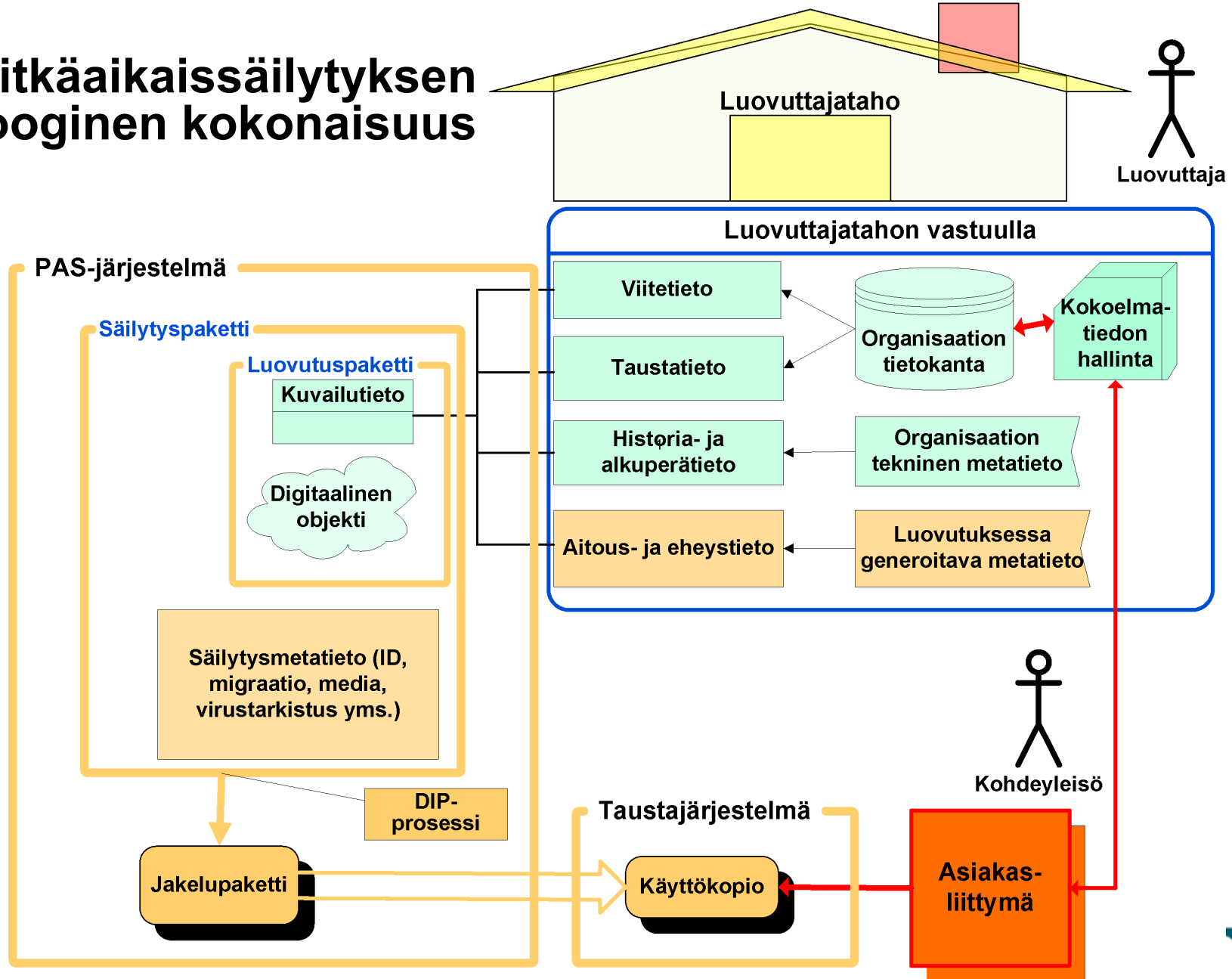


Aineisto ja asiakkaat: huomioita

- **Monenlaista aineistoa skannatuista dokumenteista multimediaan, tiedostomuodot moninaisia**
- **Määrä useita miljoonia tietueita, tavumäärän arviointi kesken**
- **Digitoinnin aste vielä pieni; raha ja työläys ongelmana**
- **Hyvin laaja kohdeyleisö**
 - Sekä yksityishenkilöitä että organisaatioita
 - Sekä Suomessa että maan rajojen ulkopuolella
- **Osa aineistoista luottamuksellisia, osa myydään, osa tilataan, osa kaikille avointa**
- **Puuttuvia tietoja kerätään lähiaikoina arkistoille, kirjastoille ja museoille lähetettävässä KDK-kyselyssä**
 - Huolellinen vastaaminen kyselyyn helpottaa jatkotyötä



Pitkäaikaissäilytyksen looginen kokonaisuus

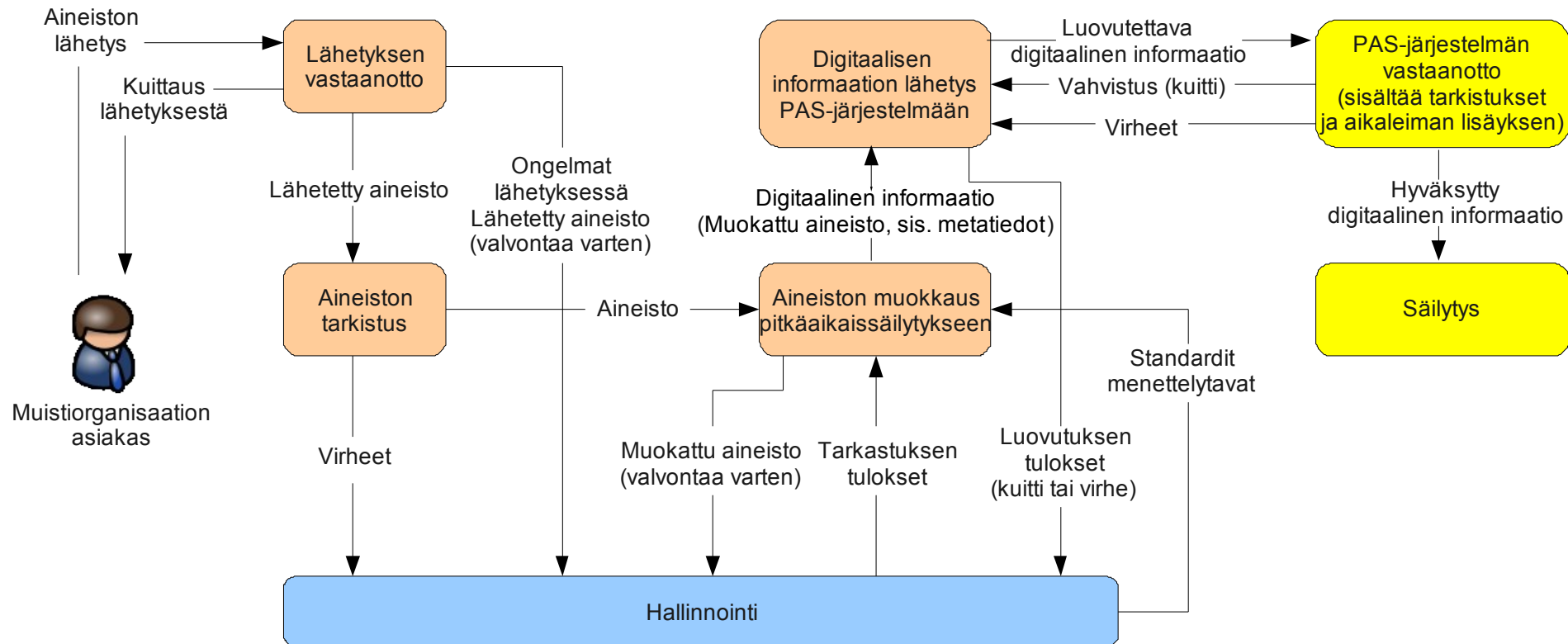


Looginen kokonaisuus

- **Luovuttajat** (arkistot, kirjastot ja museot) lisäävät aineistoon kuvailutiedot ja koostavat luovutuspaketit
 - Kuvailutieto eli luovuttajan lisäämä metatieto koostuu tulkintatiedosta, paketoitietiedosta, viitetiedosta, taustatiedosta, historia- ja alkuperätiedosta, sekä aitous- ja eheystiedosta.
- **PAS-järjestelmä** säilyttää ja hallinnoi informaatiota
 - Lisää luovutuspaketteihin säilytysmetatiedon ja muodostaa niistä säilytyspaketit. Tarpeelliset metatiedot synkronoidaan luovuttajan järjestelmiin ja asiakasliittymään.
 - Luo jakelupaketit säilytyspaketeista. Jakelupaketit (käyttökopiot) tarjotaan luovuttajalle ja asiakasliittymään.
- **Asiakasliittymä** tarjoaa sisällöt asiakkaille
 - Tarjoaa palvelut ja toiminnot PAS-järjestelmässä olevan informaation olemassaolon, kuvauksien, sijainnin ja saatavuuden selvittämiseen.
 - Mahdollistaa kokoelmarakenteiden muodostuksen



Aineiston luovutus ja vastaanotto



Käyttötapaukset

- **PAS-järjestelmän tärkeimmät käyttötapaukset ovat**
 - Digitaalisen informaation luovutus luovutuspaketteina
 - Käyttökopioiden synkronointi tuotantojärjestelmiin ja asiakasliittymään
 - Jakelupakettien pyyntö
 - Tietopaketteihin kohdistuvat hallinnolliset pyynnöt
- **Näiden lisäksi käyttötapauksia ovat mm.**
 - Hävitettäväksi esitettyjen tietopakettien poiminta, tietopakettien hävittäminen, rajoitusten tarkistus, hakemistohaku ja säilytyksen suunnitteluun liittyvät toimenpiteet
- **Käyttötapaukset on vielä syytä kartoittaa tarkemmin**

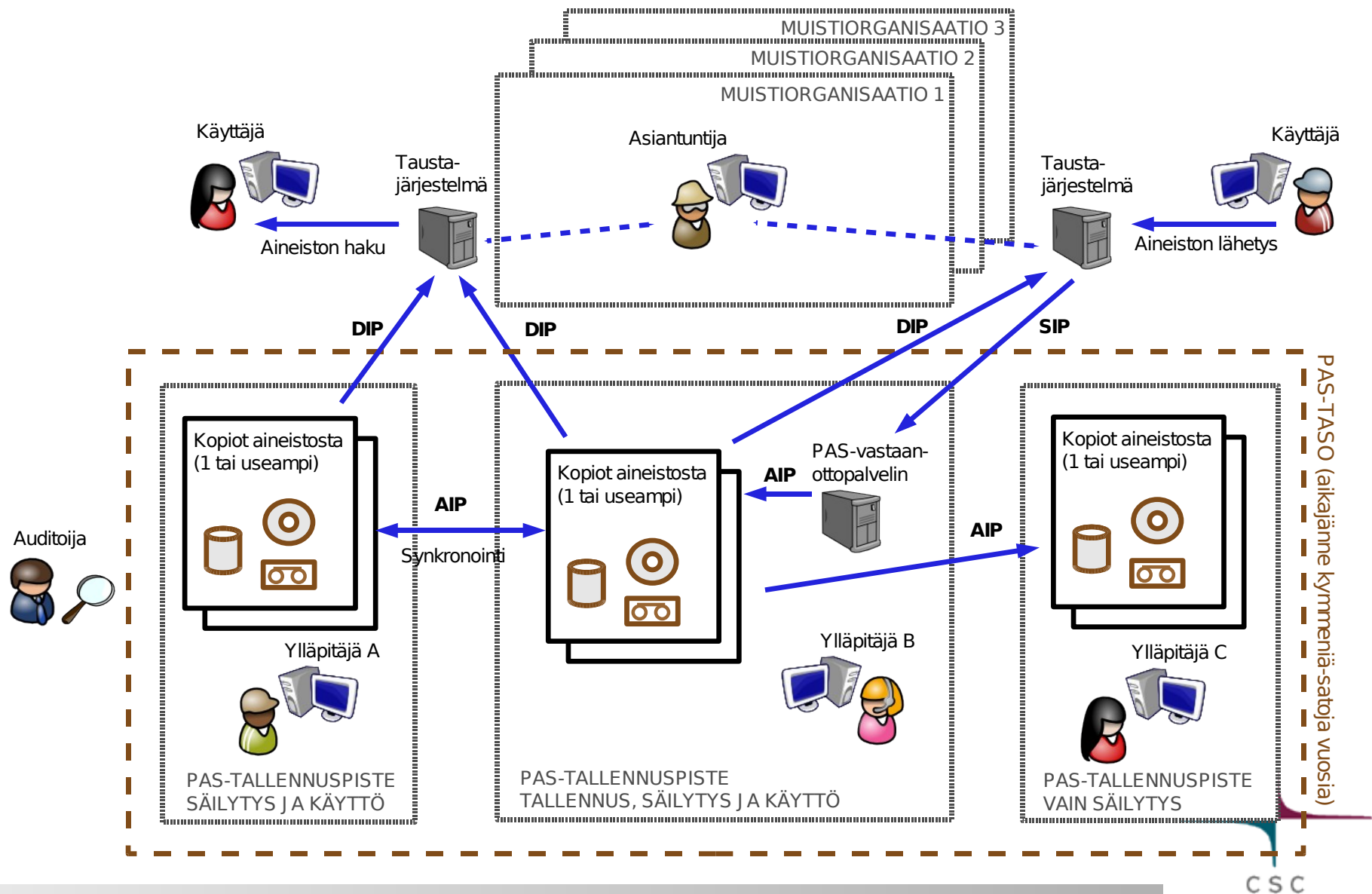


PAS-järjestelmän tekniseltä infrastruktuurilta vaadittavat ominaisuudet

- **Järjestelmässä ei saa olla yksittäistä kohtaa, jonka pettäessä koko järjestelmä pettää (single point of failure)**
- **Järjestelmän täytyy toimia, vaikka mediat, ohjelmistot ja laitteistot vaihtuvat tasaisena virtana**
 - Järjestelmän täytyy tukea erilaisuutta ja välttää lukkiutumista tiettyjen laite- tai ohjelmistotoimittajien ratkaisuihin
- **Järjestelmän täytyy jatkuvasti seurata tietopakettien eheyttä**
 - Suurinta osaa järjestelmässä olevaa aineistoa käytetään harvoin, ja se aiheuttaa suuren riskin piilevien virheiden kerääntymiseen
- **Järjestelmän tulee olla ylläpidettävissä pääosin katkotta**
- **On mietittävä, halutaanko kaiken säilytettävän digitaalisen informaation olevan saatavilla viiveettä (esim. kiintolevyillä) vai riittääkö alhaisempi taso.**



Esimerkki teknisestä infrastruktuurista



Pitkäaikaissäilytyksen riskit

- **Tekniset riskit**
 - Mm. tallennusmedian tai laitteiston rikkoutuminen, ohjelmistovirheet
- **Inhimilliset riskit**
 - Ylläpitäjän virhe
 - Hyökkäys järjestelmää vastaan (organisaation ulkoa tai sisältä)
- **Eriyisesti pitkäaikaissäilytykseen liittyvät riskit**
 - Tiedon ymmärrettävyyden häviäminen
 - Piilevät virheet ("bit rot")
- **Katastrofit**
 - Luonnonilmiöt: tulva, tulipalo ym.
 - Yhteiskunnalliset: Sotatila, rahoituksen loppuminen ym.



Riskien minimointi

- **Keinoja riskien minimoimiseksi:**
 - Useampi kuin yksi tallennusmediatyyppe (esim. kiintolevy, nauha, optiset levyt, kertakirjoitettavat mediat)
 - Automaattinen informaation eheyden tarkistus ja virheiden korjaus
 - Avoimet rajapinnat, avoimen lähdekoodin ohjelmistot
 - Turvalliset verkko- ja tilaratkaisut
 - Järjestelmän toiminnan valvonta, lokitietojen keruu ja analysointi
 - Kopioiden hajautus maantieteellisesti eri paikkoihin, eri kopioiden tallennusprosessien riittävä itsenäisyys
 - Järkevä rahankäyttö
- **PAS-organisaation on laadittava riskianalyysi ja riskienhallintasuunnitelma**
 - Riskianalyysia on myös määrääjoin päivitettävä



Standardit, lait ja normit

- **Kansainväliset standardit**
 - Open Archival Information System (OAIS) -viitemalli, joka on myös ISO-standardi (ISO 14721:2003)
 - Trustworthy Repositories Audit & Certification (TRAC)
 - Muita mm. MoReq2, METS, PREMIS, ISO-laatustandardit
- **Kotimaiset lait ja normit**
 - Julkisuuslaki, henkilö- ja tietosuojalait, valmiuslaki, tekijänoikeuslaki, laki kulttuuriaineistojen tallettamisesta ja säilyttämisestä
 - VAHTI-tietoturvaohjeet ja -määräykset
- **Muut hyvät toteutuskäytännöt, mm. tietoturvan osalta**
 - Paljon materiaalia tarjolla, kannattaa valikoida ja hyödyntää!



Lisätyötä vaativia kokonaisuuksia (erityisesti luovuttajien näkökulmasta)

- **Tuotantojärjestelmien ja PASin yhteensovitus suunnitelma**
 - Luovuttajilta tarvitaan tietoja olemassaolevista järjestelmistä
- **Käyttötapausten kerääminen ja analysointi**
- **Pitkäaikaissäilytettävän informaation määrän ja tiedostoformaattien ja metatietojen kartoitus**
 - Tästä on tulossa kysely lähiaikoina
- **Metatietojen sisällölliset ja tekniset vaatimukset**
- **Luovutukseen liittyvät toiminnot, työkalut ja rajapinnat, vaatimukset luovuttajille**
- **Vaatimukset asiakasliittymille ja rajapinnoille**
 - Sekä ihmisten käyttämät liittymät että järjestelmien rajapinnat
- **Kokonaisarkkitehtuuri, toiminnallisen määrittelyn tarkentaminen, tekninen määrittely, riskianalyysi, ...**



Lisätietoja

- **Opetusministeriön KDK-sivut**
 - http://www.minedu.fi/OPM/Kulttuuri/kulttuuripolitiikka/linjaukset_ohjelmat_ja_hankkeet/digitaalinen_kirjasto/
 - http://www.hare.vn.fi/mHankePerusSelaus.asp?h_ild=14477
- **KDK-hankkeen pitkäaikaissäilytysjaosto**
 - Puheenjohtaja Markku Nenonen, Kansallisarkisto
- **PAS-järjestelmän alustava toiminnallinen vaatimusmäärittely ja katsaus kansainvälisiin pitkäaikaissäilytys Hankkeisiin**
 - Arto Teräs ja Pirjo-Leena Forsström, CSC
- **Runsaasti kansainvälistä materiaalia verkossa, mm.**
 - Digital Preservation Tutorial (Cornell University Library):
<http://www.icpsr.umich.edu/dpm/dpm-eng/>
 - DPC/PADI What's new in digital preservation:
<http://www.dpconline.org/graphics/whatsnew/>

