Digitaalisten aineistojen pitkäaikaissäilytys - Tiedostoformaattien standardointi

Valtakunnalliset kuva-arkistopäivät
4.-5.11.2013
Vesa Hongisto

MUSEOVIRASTO

Miksi tiedostoformaatteja standardoidaan

Digitaalisen pitkäaikaissäilytyksen tasot

- Bittien säilyttäminen:
  - Alkuperäinen bittijono voidaan toistaa nykyaikaisilla laitteilla, mutta ei takeita siitä voidaanko sisältöä tulkita.
  - Säilyttämisen perusta ja optio tulevaisuuteen: antaa mahdollisuuden nostaa vaatimustasoa myöhemmin jos tarvetta ja rahaa on.
- Sisällön ymmärrettävyyden säilyttäminen:
  - Teksti, kuvat ja muu olennainen sisältö voidaan tulkita.
  - Ymmärrettävyys riippuu myös kohdeyleisöstä.
  - Käytön ja uuden tiedon tuottamisen kannalta yleensä tärkein taso.
- Alkuperäisen käyttökokemuksen säilyttäminen:
  - Kalvoesityksen tehosteet toistuvat alkuperäisessä muodossaan.
  - Vanha tietokonepeli näyttää ja kuulostaa samalta kuin 30 vuotta sitten.
  - Vaativin, mutta tutkimuksen kannalta mielenkiintoinen taso.

MUSEOVIRASTO

Miksi tiedostoformaatteja standardoidaan

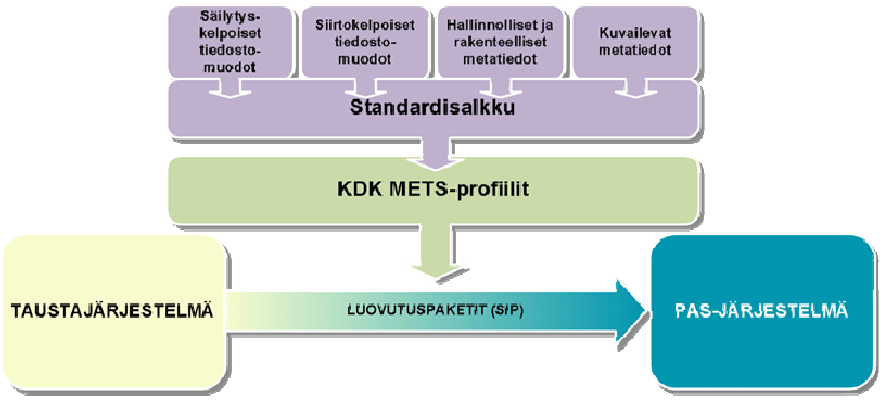
Digitaalisen pitkäaikaissäily

- Bittien säilyttäminen:
  - Alkuperäinen bittijono vo … daanko sisältöä tulkita.
  - Säilyttämisen perusta ja … asoa myöhemmin jos tarve…
- Sisällön ymmärrettävyyden säilyttäminen:
  - Teksti, kuvat ja muu olennainen sisältö voidaan tulkita.
  - Ymmärrettävyys riippuu myös kohdeyleisöstä.
  - Käytön ja uuden tiedon tuottamisen kannalta yleensä tärkein taso.
- Alkuperäisen käyttökokemuksen säilyttäminen:
  - Kalvoesityksen tehosteet toistuvat alkuperäisessä muodossaan.
  - Vanha tietokonepeli näyttää ja kuulostaa samalta kuin 30 vuotta sitten.
  - Vaativin, mutta tutkimuksen kannalta mielenkiintoinen taso.

Sisällön ymmärrettävyyden säilyttämiseksi tiedostot on muutettava kunakin ajanhetkenä käytössä oleviin formaatteihin. Tiedostoformaattien muunnossa häviää aina jonkin verran informaatiota.

MUSEOVIRASTO

HKDK:n tuottamat ohjaavat dokumentit



MUSEOVIRASTO

Säilytyssuunnitelma

- Jokaiselle pitkäaikaissäilytykseen tulevalle aineistolle tehdään säilytyssuunnitelma
- Säilytyssuunnitelmassa määritellään mitkä ovat ne ominaisuudet jotka eivät saa muuttua migraatiossa
- Jos on käytettävissä vaihtoehtoisa tedostoformaatteja valitaan se, joka parhaiten takaa halutun ominaisuuden säilymisen
- Esimerkkejä:
  - Asiakirjassa voi riittää, että informaatiosisältö, teksti, säilyy, asettelu voi muuttua
  - Taidevalokuvassa on oleellista, että myös värisävyt säilyvät

MUSEOVIRASTO

Tiedostoformaattien arviointikriteerit

- **Avoimuus**
  - Kuinka helppoa tiedostomuodosta on saada tietoja?
- **Käyttö PAS-standardina**
  - Missä määrin tiedostomuoto on muodollisesti hyväksytty pitkäaikaissäilytyksen välineeksi kansalliskirjastoissa, kansallisarkistoissa ja muissa alan laitoksissa?
- **Vakaus / yhteensopivuus**
  - (a) Missä määrin tiedostomuoto on eteen- ja taaksepäin yhteensopiva?
  - (b) Missä määrin tiedostomuoto on suojattu tiedoston korruptoitumista vastaan?
  - (c) Kuinka usein tiedostomuodosta julkaistaan korvaavia versioita?
- **Riippuvuudet / yhteentoimivuus**
  - Missä määrin tiedostomuoto on sidottu esimerkiksi tiettyyn laitteistoon tai ohjelmistoon?
- **Standardisuus**
  - Missä määrin tiedostomuoto on käynyt läpi perusteellisen standardointiprosessin?

MUSEOVIRASTO

Tiedostoformaattien arviointi

Kriteeri	Arviointiohje	Arvio
1. Avoimuus	Määritykset saatavissa yhdeltä tai useammalta seuraavista: (a) avoimen jäsenyyden järjestö [kuten W3C (World Wide Web Consortium) tai OMG (Object Management Group)] (b) Kansainvälinen standardointijärjestö (esim. ISO) (c) Tuotannonalan avoimen jäsenyyden järjestö	A
	Määritykset saatavilla vain maksusta	A€
	Määritykset mahdollisesti saatavilla useista lähteistä (ei saatu vahvistettua)	B
	Määrityksiä jakelee tai niiden jakelua valvoo yksi kaupallinen toimija tai pieni kaupallisten toimijoitten joukko.	C

MUSEOVIRASTO

Tiedostoformaattien arviointi

Kriteeri	Arviointiohje	Arvio
2. Käyttö PAS-standardina	Merkittävä osa (yli puolet) organisaatioista käyttää tai aikoo käyttää tiedostomuotoa säilytyskelpoisena tiedostomuotona	A
	Jotkut organisaatioista (alle puolet) käyttävät tai aikovat käyttää tiedostomuotoa säilytyskelpoisena tiedostomuotona	B
	Yksikään organisaatio ei käytä eikä aio käyttää tiedostomuotoa säilytyskelpoisena tiedostomuotona	C

MUSEOVIRASTO

Tiedostoformaattien arviointi

Kriteeri	Arviointiohje	Arvio
3. Vakaus / yhteen-sopivuus		
(a) Alas-/ ylöspäin yhteenso-pivuuden taso	Tiedostomuodon versio on alaspäin yhteensopiva, jos se sisältää kaikki aiemman version toiminnallisuudet Tiedostomuodon versio on ylöspäin yhteensopiva, jos sen avulla voi ongelmitta tallentaa sisältöä, joka on tarkoitettu tiedostomuodon uudemmalle versiolle (kääntäen: ohjelmisto, joka on suunniteltu tulkitsemaan tai näyttämään tiedostomuodon aiempaa versiota pystyy tulkitsemaan tai näyttämään myös tiedostomuodon nykyistä versiota).	
	Hyvä yhteensopivuus – tiedostomuoto on yhteensopiva niin ylös- kuin alaspäinkin.	A
	Kohtalainen yhteensopivuus – tiedostomuoto on vain alaspäin yhteensopiva.	B
	Huono yhteensopivuus – tiedostomuoto ei ole yhteensopiva ylös- eikä alaspäin.	C

MUSEOVIRASTO

Tiedostoformaattien arviointi

Kriteeri	Arviointiohje	Arvio
3. Vakaus / yhteen-sopivuus		
(b) Korrup-toitumisen sieto	Korruptoitumisen sieto tarkoittaa sitä, että tiedostomuoto sietää sisällön bitti- tai tavutason satunnaisia muutoksia.	
	Hyvä sietokyky – Muutokset eivät juurikaan tai lainkaan vaikuta tiedostomuodon näyttämiseen tai tulkintaan; tai tiedostomuoto sisältää menetelmiä, joilla muutokset havaitaan ja/tai korjataan	A
	Kohtalainen sietokyky – Muutokset vaikuttavat tiedostomuodon näytettävyyteen mutteivät tulkittavuuteen; tiedostomuoto voi jossakin määrin palautua muutoksista.	B
	Huono sietokyky – Kaikki muutokset vaikuttavat näytettävyyteen ja tulkittavuuteen.	C

MUSEOVIRASTO

Tiedostoformaattien arviointi

Kriteeri	Arviointiohje	Arvio
3. Vakaus / yhteen-sopivuus		
(c) Versio-päivitysten määrä	Tiedostomuodon vakaus ilmaistuna uusien versioiden tai laajennusten lukumääränä; tiedostomuodon käyttö johdannaisissa ja/tai tuotannonalakohtaisissa sovelluksissa.	
	Suuri vakaus	A
	Keskitason vakaus	B
	Epävakaus	C

MUSEOVIRASTO

Tiedostoformaattien arviointi

Kriteeri	Arviointiohje	Arvio
4. Riippumat-tomuus / yhteentoimivuus	Korkea riippumattomuus & yhteentoimivuus Korkea riippumattomuus & keskitason yhteentoimivuus Keskitason riippumattomuus & korkea yhteentoimivuus	A
	Korkea riippumattomuus & alhainen yhteentoimivuus Keskitason riippumattomuus & yhteentoimivuus Keskitason riippumattomuus & alhainen yhteentoimivuus	B
	Alhainen riippumattomuus & yhteentoimivuus Alhainen riippumattomuus & keskitason yhteentoimivuus Alhainen riippumattomuus & korkea yhteentoimivuus	C

MUSEOVIRASTO

Tiedostoformaattien arviointi

Kriteeri	Arviointiohje	Arvio
5. Standardisuus	Tiedostomuotoa sääntelee jonkin seuraavista hallitsema muodollinen prosessi: • Avoimen jäsenyyden organisaatio (W3C, OMG yms.) • Kansainvälinen standardiorganisaatio (esim. ISO) • Tuotannonalakohtainen avoimen jäsenyyden organisaatio	A
	Tiedostomuotoa sääntelee dokumentoitu prosessi, jonka on luonut yksittäinen kaupallinen toimija tai pieni kaupallisten toimijoiden joukko; tai tiedostomuotoa koskevia prosesseja ei ole dokumentoitu.	C

MUSEOVIRASTO

Tiedostoformaatit

- Säilytyskelpoiset tiedostomuodot
  - Säilytyskelpoisiksi hyväksytään sellaiset tiedostomuodot, joissa tietosisällön säilyminen ja ymmärrettävyys voidaan taata pidemmällä aikavälillä.
- Siirtokelpoiset tiedostomuodot
  - Siirtokelpoisiksi hyväksytään vain sellaisia tiedostomuotoja, joita käytetään useassa KDK:n PAS-järjestelmää hyödyntävässä organisaatiossa ja joissa pitkäaikaissäilytettävää aineistoa on runsaasti tallennettu. KDK:n PAS-ratkaisu muuntaa säilytyssuunnitelman asettamien vaatimusten ja ehtojen mukaisesti siirtokelpoisessa muodossa vastaanotetut tiedostot säilytyskelpoiseen tiedostomuotoon ennen säilyttämisen aloittamista.

MUSEOVIRASTO

Tiedostoformaatit

Sisältö	Tiedostomuoto	Avoimuus	Käyttö PAS standardina	Vakaus / yhteensopivuus			Riippumattomuus / yhteentoimivuus	Standardisuus
				Alas / ylöspäin yhteensopivuuden taso	Korruptoitumisen sieto	Versiopäivitysten määrä		
TEKSTI	Electronic Publications (EPUB)	A	B	B		A	A	A
	Extensible Hypertext Markup Language (XHTML)	A	B	B		A	A	A
	Extensible Markup Language (XML)	A	A	A		A	A	A
	Hypertext Markup Language (HTML)	A	A	B		A	A	A
	Open Document Format (ODF)	A	A	B		B	A	A
	PDF for long-term preservation (PDF/A)	A€	A	B		A	A	A
	Tekstitiedosto (plain text)	A€	A	B		A	A	A
ÄÄNI	Audio Interchange File Format (AIFF), PCMkoodattu	A	A	A		A	A	A
	Broadcast Wave Format (BWF)	A	A	A		A	A	A
	Free Lossless Audio Codec (FLAC)	A	B	A	A	A	A	A
	MPEG-4 AAC – Advanced Audio Coding (AAC)	A	B	A			A	A
	Waveform Audio Format (WAV)	A	A	A		A	A	A
ELÄVÄ KUVA	Motion JPEG 2000	A€	A	A	A	A	A	A
KUVA	Joint photographic experts group (JPEG)	A€	A			A	A	A
	Joint photographic experts group JPEG 2000 (JP2)	A€	A			A	A	A
	Tagged Image File Format (TIFF)	A	A			A	A	A
VERKKOARKISTO	Web ARChive Format (WARC)	A	A		B	A	A	A
TIETOKANNAT	Määritellään myöhemmin							

Tiedostoformaattien pakolliset metatiedot

- Aineistoja digitaaliseen pitkäaikaissäilytykseen vietäessä on tiedostoista annettava riittävän tarkka kuvaus niiden teknisistä ominaisuuksista, jotta myöhempi muunto uusiin formaatteihin on mahdollista
- KDK-hankkeessa on määritelty joidenkin aineistotyypin osalta pakolliset tekniset metatiedot
- Määrittely on tehty muutoksina suhteessa USA:n kongressin kirjaston eri tiedostotyyppien skeemoihin
- Esimerkki: tekstitiedostoissa täytyy mainita käytetty merkistö
- Teknisissä metatiedoissa on myös käytettävä samoja yksikäsitteisiä ilmaisuja esimerkiksi mittayksiköissä

MUSEOVIRASTO

Formaattikirjasto

- KDK:ssa käytetään pääsääntöisesti tiedostomuodoille PRPNOM formaattikirjaston tunnisteita
- KDK ylläpitää kontrolloitua sanastoa jossa määritetään miten kukin formaatti ja sen versio ilmaistaan

MUSEOVIRASTO

Tiedostomuotojen migraatiossa huomioitavat ominaisuudet

- Esimerkki tiedostomuun-noksessa säilyvistä, muuttuvista ja katoavista ominaisuuksista

Lähdemuoto	Kohdemuoto				Lisätiedot
GIF	PNG	JPEG	JPEG 2000	TIFF	
Datan häviötön säilyminen	X		X	X	PNG: muuttuu Deflate-muotoon, JPEG 2000: muuttuu Wavelet-muotoon, JPEG: muuttuu häviölliseen DCT-muotoon
Kuvatasot				(X)	Tarvittaessa eri tasoista voidaan tehdä useita kuvia. TIFF-muodossa ne saa kokoelmaksi.
Maksimidimensio	X	X	X	X	
Dimensio	X	X	X	X	
Esitysresoluutio (89a)	X		X	X	Konvertoituu uuteen muotoon. JPEG 2000-muodossa määritettävä kuvan fyysinen koko.
Kuvan kanavat	X	X	X	X	
Bittisyvyys	X	X	X	X	
Värikoordinaatisto	X	(X)	X	X	JPEG-muodossa konvertoitava YC_bC_r-koordinaatistoon.
Paletti	X		(X)	(X)	PNG-muodossa lokaalit paletit muuttuvat nk. ehdotetuiksi paleteiksi. JPEG 2000- ja TIFF-muodossa yhden paletin tuki.
Paletin bittisyvyys	X		X	X	TIFF-muodossa paletin bittisyvyys kaksinkertaistuu
Taustaväri	X				
Läpinäkyvyys (89a)	X		X	X	PNG-muodossa muuttuu läpinäkyvyysväriksi. JPEG 2000 ja TIFF-muodossa muuttuu kuvakanavaksi.
Lomitus	X	(X)	(X)		JPEG- ja JPEG 2000 –muodoissa muuttuu progressiiviseksi.
Kommenttikenttä (89a)	X	X	X	X	
Kuvaan upotettava teksti (89a)					
GIF-laajennusosio					
Animaatio				(X)	Animaation kuvat voidaan TIFF-muodossa koostaa kuvakokoelmaksi, mutta ilman animaation toiminnallisuutta.

Kiitos!

vesa.hongisto@nba.fi
www.nba.fi
http://www.kdk.fi/fi/pitkaaikaissailytys/maeaerittely-ja-dokumentit/5-suomi/pitkaeaikaissaeilytys/141-kdk-saeilytys-ja-siirtokelpoiset-tiedostomuodot

MUSEOVIRASTO