Digitaalisen kirjaston metadatat

Juha Hakala

Laadulliset muutokset ovat yleensä määrällisiä ongelmallisempia, niin työssä kuin elämässä yleensäkin. Siinä missä määrällisiä muutoksia voidaan usein ennakoida, laadulliset tulevat yleensä yllättäen. Ja laadullisiin muutoksiin ei läheskään aina voida vastata perinteisin keinoin.

Digitaalinen kirjasto on laadullinen muutos. Kirjoituksessani tarkastelen digitaalista kirjastoa metadatan eli resurssien kuvailun näkökulmasta. Muutoksia voi luonnehtia suuriksi.

Mikä digitaalinen kirjasto?

Jotta lukija tietää, mistä keskustellaan, on syytä aluksi määritellä digitaalinen kirjasto. Aina avuliaas Wikipedia valistaa, että englanninkielinen termi on otettu käyttöön 1988, ja määrittelee sen näin (katso http://en.wikipedia.org/wiki/Digital_library):

> *A digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored locally, or accessed remotely via computer networks. A digital library is a type of information retrieval system.*

Suomeksi digitaalinen kirjasto on pari vuotta sitten määritelty näin (katso http://www.kansalliskirjasto.fi/kirjastoala/tietolinja/0207/pk0207.html):

> *Digitaalinen kirjasto edistää tiedon ja kulttuuriperinnön saatavuutta verkossa. Digitaalinen kirjasto tuottaa ja kerää, hallinnoi sekä säilyttää digitaalisia sisältöjä ja tarjoaa niistä palveluita asiakaslähtöisesti, suunnitelmallisesti ja laadultaan mitattavasti.*

Pohja määritystyölle oli DELOS-projektissa laadittu määritelmä, jonka mukaan digitaalinen kirjasto voi olla virtuaalinen organisaatio. Näin on asian laita esimerkiksi Kansallisessa digitaalisessa kirjastossa (http://www.kdk2011.fi/), jossa on mukana 35 muistiorganisaatiota.

Projektin tavoitteena on kirjastojen, arkistojen ja museoiden sähköisten aineistojen saatavuuden parantaminen ja pitkäaikaissäilytys. Tavoitteena on rakentaa yhteinen asiakasliittymä sekä digitaalinen arkisto. Pitkäaikaisuudella tarkoitetaan tässä vähintään kymmenien ja mahdollisesti satojen vuosien ajanjaksoa, jonka mittaan laitteistot ja ohjelmistot ennättävät mennä moneen kertaan uusiksi.

Tässä artikkelissa digitaalisella kirjastolla tarkoitetaan erityisesti sitä toiminnallista kokonaisuutta, jonka opetusministeriön koordinoima Kansallisen digitaalisen kirjaston hanke tulee rakentamaan. Erilaisia digitaalinen kirjasto –sateenvarjon alle soveltuvia järjestelmiä on Suomessakin rakennettu jo 90-luvulta lähtien, mutta näissä palveluissa ei välttämättä ollut yhteisiä nimittäjiä edes yhden organisaation sisällä. Vasta 2000-luvun lopulla digitaalisen kirjaston kokonaiskuva ja tekninen pohja on kirkastunut niin, että tiedämme suhteellisen tarkasti, mistä me puhumme kun puhumme digitaalisesta kirjastosta.

Metadatan muutostarpeet

Perinteisen aineiston kuvailu perinteisessä kirjastossa on vakiintunutta toimintaa. Sen pohjana on luettelointisääntöjen ja MARC-formaatin muodostama kokonaisuus, joka on palvellut meitä hyvin: ilman näitä välineitä emme pystyisi hankkimaan ulkomaisia kirjastojärjestelmiä, emmekä poimimaan tietueita lähes kaikkialta maailmasta.

Viimeisen kymmenen vuoden aikana AACR-MARC -paradigmaan on syntynyt säröjä. Nykyinen toimintamalli ei kaikin osin vastaa digitaalisen kirjaston vaatimuksia, vaikka se perinteisellä sovellusalueellaan on yhä toimiva.

Osittain kyse on myös maailmankuvan muutoksesta. FRBR-malli ja sen varaan rakentuvat Resource Description and Access -kuvailusäännöt ottavat nykyistä paremmin huomioon sen, että meillä on kokoelmissamme teoksia, joita edustaa kasvava määrä erilaisia painettuja ja digitaalisia manifestaatioita. Siirtymä perinteisistä kuvailusäännöistä RDA:han on yhä kesken. RDA:ta tukevia sovelluksia ei ole, mutta asiantuntijat ovat jo pääosin siellä, minne me muut olemme vasta menossa, FRBR/RDA-pohjaisessa maailmassa, jossa myös teosten eikä vain niiden manifestaatioiden kuvailulla on keskeinen merkitys.

Käytännön tasolla toinen selvästi näkyvä muutos on se, että digitaalisessa kirjastossa tarvitaan aivan uudenlaista metadataa, jotta sähköisiä aineistoja kyetään tyydyttävästi hallinnoimaan. MARC-formaatti, vaikka siihen onkin vuosien mittaan lisätty vähittäin elektronisen aineiston kuvailussa tarvittavia tietoelementtejä, ei sisällä läheskään kaikkea sähköisten aineistojen kannalta tarpeellista tietoa.

Metadatan lajit digitaalisessa kirjastossa

Erinomaisessa joskin jo hieman vanhentuneessa katsauksessaan Metadata for digital libraries: State of the art and future directions (katso http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf) Richard Gartner jaottelee digitaalisen kirjaston metadatan kolmeen ryhmään:

1. Kuvaileva metadata kattaa jotakuinkin kaiken sen, mikä kuuluu perinteisen kuvailun alaan. Luetteloinnin kohteena on dokumentin intellektuaalinen sisältö, ja tavoitteena on edistää dokumentin haettavuutta ja helpottaa esimerkiksi sen relevanssin arviointia.
2. Hallinnollisen metadatan tehtävänä on mahdollistaa muun muassa sähköisten dokumenttien jakelu ja pysyvä säilyttäminen. Tämä ryhmä voidaan jakaa edelleen
   a. Tekniseen metadataan, joka liittyy dokumentin pitkäaikaissäilytykseen ja muuhun käsittelyyn;
   b. Käyttöoikeustietoihin, joissa kuvataan dokumenttiin liittyvät tekijänoikeudet sekä muut käyttöä koskevat rajaukset; sekä
   c. Pitkäaikaissäilytyksen metadataan, joka kuvaa dokumentin luontiin ja myöhempään käsittelyyn (migraatiot) liittyviä tapahtumia ja toimijoita.
3. Rakenteellinen metadata kuvaa dokumentin fyysisen rakenteen siten, että dokumentti kyetään esittämään järkevästi (esimerkiksi kirjan kuvatiedostoiksi digitoidut sivut oikeassa järjestyksessä). Lisäksi on voitava esittää dokumentin looginen rakenne (esimerkiksi kausijulkaisun numerossa olevat artikkelit), jotta haku dokumentista ja sen esittäminen saadaan tehokkaiksi.

Tarve tuottaa hallinnollista ja rakenteellista metadataa ovat osa laadullista muutosta siirryttäessä perinteisestä kirjastosta digitaaliseen kirjastoon ja perinteisistä painetuista aineistoista sähköiseen aineistoon. Sopeutuminen tähän muutokseen edellyttää kirjastoista uudenlaista osaamista ja välineistöä. Joudumme miettimään myös toimintamme prioriteetteja ja yhteistyömahdollisuuksia: eritoten pitkäaikaissäilytyksen haaste on niin suuri, ettei mikään organisaatio selviä siitä yksin.

Perinteinen kuvaileva metadata on tiukasti sidoksissa organisaatioiden työprosesseihin ja tuotantojärjestelmiin. Sen vuoksi kirjastoissa tuotettu kuvailu poikkeaa museoiden tai arkistojen

tuottamasta. Säännöt, formaatit sekä osin kuvailun periaatteet ja tavoitteet ovat erilaisia.

Eri muistiorganisaatioissa tuotetulla kuvailevalla metadatalla voi ja pitää silti olla yhteinen semanttinen ydin – KDK-hankkeen asiakasliittymähankkeen menestys on osittain riippuvainen tästä yhteismitallisuudesta, jota Suomessa on pyritty avittamaan Kamut-hankkeilla.

Hallinnollinen ja rakenteellinen metadata eivät nykyisen käsityksen mukaan ole organisaatioista riippuvaisia. Kuvatiedoston tekniseen metadataan vaikuttaa ensisijaisesti se, millaisella skannerilla ja parametreilla kuva on skannattu, eikä se, onko skannaus tehty kirjastossa, arkistossa vai museossa. Samoin Word-tiedoston migraatiossa syntyvä metadata riippuu esimerkiksi käytettävistä välineistä ja migraation ambitiotasosta, ei siitä missä organisaatiossa työ tehdään.

Rakenteellinen metadata

Painetun tekstidokumentin rakennetta ei ole ollut tarpeen kuvata tarkasti kirjastojen tietojärjestelmissä. Ihminen ymmärtää kausijulkaisun tai monografian rakenteen ilman tietojärjestelmän antamaa tukea. Korkeintaan kuvailuun on lisätty huomautuksia esimerkiksi siitä, että kirjassa on kuvitus tai vaikkapa bibliografia. Rakenteen esittämiseen koneymmärrettävässä ”rautalankamuodossa” ei puolestaan ole edellytyksiä ennen kuin aineisto on digitaalista.

Digitaalisessa kirjastossa dokumentin fyysistä ja loogista rakennetta koskevat tiedot ovat oleellisia, koska niiden avulla aineisto voidaan esittää ja käsitellä tehokkaasti. Fyysisen rakenteen erittelevän metadatan avulla voimme esimerkiksi kertoa, että digitoitu kirja koostuu 300 sivutiedostosta, ja lisäksi voidaan kertoa näiden tiedostojen esittämisjärjestys. Loogisen rakenteen kuvauksessa eritellään halutulla tarkkuustasolla ao. kirjan jakautuminen osiin (esimerkiksi nimiösivu, esipuhe, luvut, hakemisto).

Digitoinnissa ja muussa digitaalisessa julkaisemisessa rakenteellisella metadatalla on iso vaikutus esimerkiksi aineiston haettavuuteen ja käytettävyyteen. Rakenteistamisen tasosta riippuen asiakkaalle voidaan esittää esimerkiksi vain sanomalehden koko numero. Kun rakenteistaminen viedään pidemmälle, voidaan asiakkaille toimittaa yksittäisiä artikkeleita tai mahdollisesti vain niihin sisältyviä osakokonaisuuksia, kuten kuvia.

Vastaavasti haku voidaan kohdistaa yksinkertaisimmillaan vain kokotekstiin, mutta perusteellisesti rakenteistettua dokumenttia voidaan etsiä artikkeleiden nimillä, tekijöillä, artikkeleissa mainituilla henkilöiden nimillä ja niin edelleen.

Mikä on oikea rakenteistamisen taso? Tähän ei ole valmista eikä yleispätevää ratkaisua. Ambitiotaso riippuu käytettävissä olevista resursseista ja tulostavoitteista. Muutamissa projekteissa – esimerkiksi Norjan kansalliskirjastossa – on päädytty vaatimattomaan tasoon, jotta aineistoa saadaan tarjolle paljon ja nopeasti. Kun tekniset perusratkaisut on tehty oikein, aineistoa voidaan myöhemmin, resurssien sen salliessa, rakenteistaa pidemmälle.

Kansalliskirjastossa ei haluttu mennä suoraan Norjan malliin, mutta toisaalta meillä ei ole resursseja myöskään kattavaan rakenteistamiseen. Kattavan sisäisen keskustelun jälkeen massadigitointihankkeessamme on määritelty tavoitetasot monografioille ja kausijulkaisuille. Esimerkiksi novellikokoelmien novellit ja kausijulkaisun artikkelit kuvataan, mutta ei vaikkapa runokokoelmien yksittäisiä runoja. Tietyn runoilijan töiden kriittistä editiota tuottava projekti voi joskus tulevaisuudessa tehdä toisenlaisen linjauksen tarpeen mukaan; Kansalliskirjaston käyttämä docWORKS-sovellus mahdollistaa tämän. Käyttäjien kannalta kattava rakenteistaminen on etu - he tuskin panisivat pahakseen sitä, että löytävät haluamansa runon sen nimen tai ensimmäisen säkeen avulla, ja voisivat luoda pysyviä linkkejä suoraan kyseiseen teokseen.

Rakenteellisen metadatan roolia tulevissa digitaalisten aineistojen hakujärjestelmissä voi olla vaikea hahmottaa. Perinteisessä kirjastossa tämäntyyppistä metadataa ei ole liiemmälti tuotettu, eikä sillä ole ollut merkittävää roolia. Toisaal-

ta aivan viime aikoihin asti digitaalisen kirjaston hankkeissa on tuotettu rakenteellista metadataa vain rajoitetusti - osittain sen vuoksi että ilman tarkoitukseen soveltuvaa standardia ja ohjelmistotyökalua tämäntyyppisen metadatan tuottaminen edes ns. karvalakkimallilla vie liikaa aikaa.

Käyttöoikeustiedot

Perinteisessä kirjastossa käyttöoikeuksien hallinta ei ole ollut erityinen ongelma. Nidetietueisiin tallennetaan lainattavuustieto, ja asia on sillä selvä: kirja joko voidaan lainata tai sitten ei, ja samat pelisäännöt ovat voimassa kaikille ajasta ja paikasta riippumatta.

Sama digitaalisen dokumentin kopio voi olla samaan aikaan joko asiakkaan saatavilla tai ulottumattomissa, riippuen käyttäjästä ja/tai hänen sijainnistaan. Lisäksi termi ”käyttö” on sähköisillä aineistoilla monitulkintainen; yksinkertaisimmillaan se voidaan tulkita aineiston lukuoikeudeksi, mutta se voi merkitä myös oikeutta kopioida aineisto omaan käyttöön tai jopa muuntaa sen sisältöä digitaalisessa kirjastossa. Viimemainitun oikeuden tarvitsee pitkäaikaissäilytysjärjestelmän henkilökunta silloin, kun dokumentti muunnetaan tiedostomuodosta toiseen käytettävyyden takaamiseksi.

Käyttöoikeudet voivat muuttua nopeastikin esimerkiksi sopimuslisenssien muutosten myötä. Toisaalta esimerkiksi elektronisten vapaakappaleiden käyttöoikeudet säilynevät muuttumattomina pidempään kuin yhdenkään atk-järjestelmän elinkaari. Sähköisen aineiston käyttöoikeuksia koskeva metadata voi olla monimutkaista, mutta se on silti kyettävä siirtämään sovelluksesta toiseen siten, että sen koneluettavuus säilyy. Tähän tavoitteeseen voidaan päästä vain soveltamalla yhteisiä kuvailuperiaatteita ja formaatteja.

Tekninen metadata

Teknisen metadatan tallentamisessa ei sinänsä ole mitään uutta. Sähköisiä aineistoja kuvaillessaan muistiorganisaatiot ovat tallentaneet tätäkin informaatiota. Tavoitteena on kuitenkin ollut ensi sijassa aineiston haettavuuden ja käytettävyyden edistäminen, ei sen autenttisuuden ja pitkäaikaissäilytyksen takaaminen tai sen mahdollistaminen, että esimerkiksi kuvatiedosto pystytään vielä vuosikymmenienkin jälkeen esittämään oikein.

Pitkäaikaissäilytys on prosessi, joka alkaa silloin kun dokumentti luodaan, ja päättyy kun aineisto tuhotaan tai tuhoutuu omia aikojaan. Tekninen metadata on tästä erityisen hyvä esimerkki. Jotta kuva voitaisiin esittää virheettömästi, sitä luotaessa esimerkiksi skannaamalla on otettava talteen paljon tietoa itse skannaustapahtumasta, jotta data voidaan jälkikäteen tulkita oikein. Lisäksi kuvadatasta on laskettava mahdollisimman nopeasti tarkistussumma, jotta datan autenttisuus on myöhemmin varmistettavissa. Jos digitaaliseen arkistoon tulee datan luovutuspaketti vailla tarkistussummaa, arkisto ei pysty tarkistamaan, onko aineisto säilynyt siirron ajan muuttumattomana.

Rakenteellisen metadatan tavoin tekninen metadata luodaan pääosin ohjelmallisesti, ja työhön tarvitaan tarkoitukseen soveltuvia ohjelmistoja sekä tiedostomuotokohtaisia formaatteja, joista lisää tulevissa luvuissa.

Tekniseen metadataan liittyy toiminto joka tunnetaan kansainvälisesti nimellä format library; suomenkielinen käännös on horjuen joko formaatti- tai tiedostomuotokirjasto. Kyse on järjestelmästä jossa kuvaillaan ”riittävän tarkasti”

- olemassa olevat tiedostomuodot – esimerkiksi kaikki keskeiset kuva-, ääni- ja tekstiformaatit,
- sovellukset joiden avulla nämä tiedostomuodot ovat käytettävissä, sekä
- mahdollisesti myös tietoa siitä, mitä tapahtuu kun tiedostomuotoja konvertoidaan näillä sovelluksilla uudempiin.

Kun Fennica-tietueessa kerrotaan, että jonkin tiedoston muoto on PDF, lähes jokainen tämän päivän Internet-käyttäjä tietää että tätä tiedostomuotoa voi katsella esimerkiksi Acrobat-sovelluksella. Hieman harvempi osaa nimetä sovelluksia, joiden avulla PDF-tiedostoja voi muokata.

Mutta 50 tai 100 vuoden kuluttua on jo vaikeata löytää sen paremmin asiakkaita kuin sovelluksiakaan joille PDF on tuttu. ”Kello lyö – kaikki” sanoi Stanislaw Jerzy Lec, ja tämä pätee myös sähköiseen aineistoon, ellei sitä pidetä ajan tasalla, ja tässä tarvitaan formaattikirjaston apua.

Pitkäaikaissäilytyksen metadata

Formaattikirjastojen sisältämät tiedot ovat pitkäaikaissäilytyksen perusta. Valitettavasti näiden tietojen kokoaminen on raskasta, yksikään nykyisistä palveluista ei ole kattava. Hyvä esimerkki on Iso-Britannian Kansallisarkiston PRONOM-palvelu (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx), jossa on ongelmia sekä kuvailtujen tiedostomuotojen että kuvailun katteen ja ajantasaisuuden osalta. Tästä huolimatta PRONOM lienee tunnetuin formaattikirjasto.

Tavoitteena on, että formaattikirjastojen metatiedoista tulee samalla tapaa vaihdettavia kuin tätä nykyä MARC-tietueista. Tarve näiden tietojen vaihtamiseen on suuri. Muuten digitaaliset arkistot joutuvat tekemään valtavan määrän päällekkäistyötä ja pahimmassa tapauksessa dataa menetetään tekemällä virheitä, jotka joku muu on jo kantapään kautta löytänyt ja dokumentoinut aiemmin. Koska yhteistyö on ainoa järkevä toimintamalli, formaattikirjastot kypsynevät palveluina sitten kun pitkäaikaissäilytyksestä tulee rutiinitoimintaa. Tähän menee esimerkiksi Suomessa vielä muutamia vuosia, maailmanlaajuisesti todennäköisesti vielä paljon pidempään.

Mitä onkaan varsinainen pitkäaikaissäilytyksen metadata, jos tiedostojen tekniset ominaisuudet on jo kuvattu toisaalla? Tavoitteena on kuvata yksinkertaisesti se, mitä PAS-sovelluksessa oleville aineistoille on tehty sekä se, kuka/mikä tästä operaatiosta vastaa. Voidaan esimerkiksi kertoa, että Word 97 -muodossa ollut tekstitiedosto on konvertoitu Office 2007 -paketin Word-sovelluksella OOXML-muotoon, työn teki N.N. 25. päivä kesäkuuta 2013 ja että tässä yhteydessä tiedoston intellektuaalinen sisältö on muuttunut samana, mutta yksi tiedostoon liittynyt makro on menetetty. Tätä tietoa voidaan käyttää hyväksi myöhemmissä konversioissa varsinkin, jos käytetyissä työkaluissa tai työmenetelmissä on ollut systemaattista vikaa.

Pitkäaikaissäilytyksen metadata, etenkin jos laskemme mukaan formaattikirjastojen sisältämän tiedon, on monimutkainen kokonaisuus, jolla ei haluta raskauttaa sen paremmin kirjastojen tuotantojärjestelmiä kuin niiden käyttäjiäkään. Tämä metadata syntyy ja sitä ylläpidetään sähköisen aineiston pitkäaikaissäilytykseen tarkoitetussa sovelluksessa. Kun aineistot esimerkiksi migraation jälkeen lähetetään takaisin taustajärjestelmiin asiakaskäyttöä varten, pitkäaikaissäilytykseen liittyviä tietoja ja esimerkiksi dokumenttia alkuperäisessä muodossaan ei yleensä liitetä jakelupakettiin, joka PAS-sovelluksesta lähtee ulkomaailmaan.

Pitkäaikaissäilytykseen tarvitaan monimutkainen sovellus, joka on rakennettu nimenomaan tätä tarkoitusta varten. Lisäksi tarvitaan teknisten asiantuntijoiden joukko huolehtimaan järjestelmän teknisestä infrastruktuurista, sekä sisältöasiantuntijoiden verkosto, joka huolehtii aineistojen migraatioiden suunnittelusta ja toteutuksesta.

Eriyttämällä PAS-toiminta muistiorganisaatioiden tuotantojärjestelmistä helpotamme oleellisesti julkaisuarkistojen ja muiden e-aineistojen hallintaan tarkoitettujen sovellusten kehittämistä ja käyttöä. Eriyttäminen edellyttää selkeät rajapinnat, joiden kautta metadata ja dokumentit siirtyvät tuotantojärjestelmien ja PAS-sovelluksen välillä. Tämä on kuitenkin pieni hinta siitä, ettei meidän tarvitse luoda yhtä järjestelmää, jonka pitäisi huolehtia kaikesta. Tämä ei tietenkään sulje pois sitä, että luodaan modulaarinen kokonaisjärjestelmä, jonka eri palat kommunikoivat keskenään niin tiiviisti, että käyttäjälle syntyy mielikuva yhdestä sovelluksesta.

Metadataformaatit ja container-standardit

Metadatan esittäminen koneymmärrettävässä muodossa edellyttää formaatin jossa data voi-

daan esittää rakenteisesti, nykyään tyypillisesti XML-muodossa.

Perinteisessä kirjastossa MARC21-yhtenäisformaatti ja integroitu kirjastojärjestelmä riitti lähes kaikkeen. Digitaalisessa kirjastossa tarvitsemme sekä useita formaatteja, että joukon sovelluksia, joita käytetään rinnan kattavan palvelukokonaisuuden luomiseksi.

Ennen kuin sukellamme metadataformaattien pariin, on syytä kertoa että myös digitaalisen kirjaston metadatalle tarvitaan vaihtomuoto. Eikä ainoastaan metadatalle: myös dokumentit, eli tiedostot joista ne rakentuvat, on kerättävä samaan pakettiin tai ainakin linkitettävä siihen. Vasta tällöin datan siirtäminen sovelluksesta toiseen – esimerkiksi taustajärjestelmästä PAS-sovellukseen ja takaisin – on toteutettavissa.

Näille standardeille, joita englanniksi kuvataan sanalla container eli kontti – ei ole vakiintunutta suomalaista nimeä. Yksi mahdollisuus on perinteinen vaihtomuotostandardi, joka voi kuitenkin ohjata ajattelua turhan kapealle uralle. Myös paketointi- ja konttistandardi-termit ovat vilahdelleet asiaa koskeneessa keskustelussa, jossa ei ole vielä saavutettu kompromissia.

Yleiskäyttöiseksi digitaalisen kirjaston vaihtomuotostandardiksi on kaksi vaihtoehtoa: METS eli Metadata Encoding and Transmission Standard (http://www.loc.gov/standards/mets/) ja MPEG21 DIDL eli Digital Item Declaration Language (http://xml.coverpages.org/mpeg21-didl.html). Lähes kaikki digitaalisen kirjaston hankkeet soveltavat edellistä.

Konversio on rakennettu ainakin METSistä DIDL-muotoon. Ei ole kuitenkaan varmaa voidaanko sama data palauttaa häviöttömästi edelleen METSiin. Asiantuntijoiden mukaan DIDL:n ongelmana on se, dokumentin loogisen ja fyysisen rakenteen kuvaus ovat kytköksissä toisiinsa, mikä voi tehdä DIDL-dokumenteista hyvin monimutkaisia. METSissä nämä kaksi rakennetta voidaan esittää toisistaan riippumatta.

KDK-hankkeessa METS on vahva kandidaatti paitsi vaihtomuotostandardiksi, myös pitkäaikaissäilytysjärjestelmän säilytyspaketin rakenneratkaisuksi. Päätös asiasta pyritään tekemään syksyllä 2009. Jos projekti sitoutuu METSiin, tarvitaan vielä sovellusohje eli profiili, jossa täsmennetään esimerkiksi sitä, mitä metatietoja säilytyspakettien tulee sisältää.

Rakenteellisen metadatan standardit

Kuten edellisestä luvusta kävi ilmi, rakenteellisen metadatan esittämiseen sovelletaan digitaalisissa kirjastoissa yleensä METS-formaattia. METSiä on käytetty jo noin viiden vuoden ajan, ja tänä aikana on tuotettu useita miljoonia METS-paketteja. Tähän nähden METS-muotoisen datan esittämiseen tarkoitettuja sovelluksia on rakennettu suhteellisen vähän - hyvin toimivia ohjelmistoja on markkinoilla vain muutamia.

METSin ohella on luotu erikoistarkoituksiin pari muuta standardia. WARC eli Web Archive File Format on tarkoitettu verkkoarkistoihin kerätyn datan säilyttämiseen, ja toimii tässä rajatussa tehtävässä erinomaisesti. MXF (Material Exchange Format) on puolestaan METSin kaltainen määritys jonka alaa on liikkuva kuva.

Kuvailevan metadatan standardit

KDK-hankkeessa luodaan muistiorganisaatioiden yhteinen asiakasliittymä, johon indeksoidaan metadataa satojen kirjastojen, arkistojen ja museoiden viitetietokannoista.

Näillä organisaatioilla on käytössään 16 formaattia, joista osa voidaan jättää huomiotta sen vuoksi että data konvertoidaan muuhun muotoon ennen sen lähettämistä asiakasliittymään. KDK-hankkeen standardisalkkuun ovat päätymässä seuraavat työkalut:

- MARC21
- Dublin Core
- MODS
- EAD (Encoded Archival Description)
- EAC (Encoded Archival Context)
- CDWA (Categories for the Description of Works of Art)

- CIDOC-CRM (CIDOC Conceptual Reference Model)
- SPECTRUM (Standard ProCedures for CollecTions Recording Used in Museums
- VRA Core (Visual Resources Association)
- Film identification – Minimum set of metadata for cinematographic works (EN 15744)

Asiakasliittymään toimitettava metadata tulee normalisoida ennen kuin se indeksoidaan. Tämä tarkoittaa tiedon harmonisointia – esimerkiksi päivämäärien esittämistapa pitää yhtenäistää – mutta myös ja ennen kaikkea tietoelementtien ”mäppäystä”; sen arvioimista, mitkä kentät eri formaateissa vastaavat toisiaan. Vasta tällöin eri formaateissa saapuva data voidaan tallentaa asiakasliittymäsovelluksen sisäisessä muodossa.

Vastaavia hankkeita on vireillä muuallakin, hyvä esimerkki on DOI-yhteisön Vocabulary Mapping Framework (katso http://www.doi.org/news/VMF_project_announcement_090615.pdf).

Osa harmonisoinnista voidaan ja pitää tehdä asiakasliittymäsovelluksessa. Joissakin tapauksissa jälkikäteen tapahtuva harmonisointi ei onnistu. Esimerkki tästä on auktoriteettikontrolli: Kansalliskirjaston Kustaa Mauri Armfelt sekä Gustaf Mauritz Armfelt ja Kansallisarkiston Gustav Mauritz Armfelt eivät lyö asiakasliittymässä veljen kättä keskenään – ellei sovellukselle erikseen kerrota että kyse on samasta henkilöstä. Myös sisällönkuvailun jälkikäteinen harmonisointi on vaikeaa, varsinkin jos organisaatiot eivät edes käytä samaa asiasanastoa/ontologiaa tai luokitusta. Näissä tapauksissa harmonisointi on tehtävä jo kuvailuvaiheessa, ja edellyttää ensisijaisesti poliittisia päätöksiä ao. organisaatioissa.

Käyttöoikeustiedot

Käyttöoikeustietojen ilmaisemiseen on olemassa ainakin neljä formaattia:

- METS Schema for rights declaration
- XrML (eXtensible Rights Markup Language)
- ODRL (Open Digital Rights Language)
- PREMIS Rights

Mikään näistä ei ole itsestään selvä markkinajohtaja, ja useimpia vaivaa tiivis suhde amerikkalaiseen lainsäädäntöön.

KDK-hankkeessa pidetään kansainvälisten esimerkkien tarkastelun ja sisäisten keskustelujen jälkeen varteenotettavana vaihtoehtona PREMIS-formaatin Rights-osion käyttöä. Tällöin käytännössä ratkaisu olisi, että normaalitapauksissa viitteeseen tallennetaan vain linkki verkossa olevaan tiedostoon, jossa kuvataan ao. ryhmään kuten e-vapaakappaleisiin kuuluvien aineistojen käyttöoikeudet. Tämä on ylläpidollisesti kevyt ja kohtuullisen joustava toimintamalli.

Käyttöoikeustietojen ilmaisemisen haaste on se, että dokumentin eri osat voivat olla eri asemassa. Kotimaisen sanomalehtiartikkelin teksti ja siinä oleva kansainvälisen kuvatoimiston kuva eivät ole niin sanotusti samalla viivalla. Tämä on konkreetti esimerkki siitä, että digitaalisessa kirjastossa kuvailu on esimerkiksi tunnisteiden tai käyttöoikeuksien osalta toisinaan pakko viedä osakohdetasolle.

Teknisen metadatan standardit

Tekninen metadata on se osa-alue, jossa formaattien kehittäminen on eniten kesken. Osittain tämä johtuu tehtävän haasteellisuudesta. On käytännössä mahdotonta kehittää kaikenkattavaa formaattia, minkä vuoksi standardointi aloitettiin still-kuvista, minkä jälkeen on hiljalleen tehty muita määrityksiä. Valmiita standardeja on silti vain kaksi:

- MIX (NISO Metadata for Images in XML Schema, http://www.loc.gov/standards/mix//)
- textMD (Technical Metadata for Text, http://www.loc.gov/standards/textMD/)

Osasyy kehittämistyön hitauteen on teknisen metadatan standardien monimutkaisuus. Kun tavoitteena on mahdollistaa kaiken still-kuvaan liittyvän teknisen tiedon tallentaminen, lopputulos on maallikolle mahdoton ymmärtää. MIX-

standardin versio 2.0 on tästä varsin vakuuttava esimerkki (katso http://www.loc.gov/standards/mix/mix20/mix20.xsd). Sitä vilkaistuaan voi pohtia myös sitä, miten pitkälle MARC21-formaatin rahkeet riittävät teknisen metadatan tallennuksessa silloin, kun tavoite on aineiston pitkäaikaissäilytyksen takaaminen.

Videolle ja audiolle, puhumattakaan harvinaisemmista aineistoista, ei ole olemassa valmista standardia, vaan parhaimmillaankin vain kokeiluluontoisia määrityksiä jotka voivat olla pahasti vanhentuneita. Yksi esimerkki tästä on AudioMD (http://www.loc.gov/rr/mopic/avprot/DD_AMD.html). Se on kehitetty vuonna 2003, eikä määritystä ole sen jälkeen pidetty yllä. Silti monet projektit soveltavat AudioMD:tä, koska parempaakaan vaihtoehtoa ei ole.

Digitaalisen kirjaston rakennuspuuhissa tekninen metadata on yksi ongelmallisista alueista: standardeja puuttuu, ja silloinkin kun se on olemassa, tietojen tallentaminen on vaikeaa jos sovellukset eivät sitä tue.

Pitkäaikaissäilytyksen metadatan standardi

Vielä noin 10 vuotta sitten pitkäaikaissäilytyksen metadatan standardointi oli täysin kesken. Ne organisaatiot, jotka käynnistivät oman PAS-hankkeensa jo tuolloin, joutuivat keksimään formaatin itse, ja sen jälkeen etsimään myötämielisen ohjelmistotoimittajan. Asiaa ei helpottanut se, että formaattien välillä oli suuria eroja, koska tulokulma pitkäaikaissäilytykseen vaihteli. Esimerkiksi EU:n rahoittamassa NEDLIB-hankkeessa luotiin emulointia tukeva formaatti, kun useimmat muut ottivat pitkäaikaissäilytyksen strategiaksi migraation.

Tein 90-luvun lopulla omaan käyttööni vertailutaulukon pitkäaikaissäilytyksen formaateista. Tulos oli sikäli lohduton, että pohja yhteisymmärrykselle vaikutti melko heikolta. Jos kaikki formaatit olisi koottu yhteen, lopputulos olisi ollut sekä sekava että vaikeakäyttöinen.

Tässä valossa PREMIS (PREMIS Data Dictionary for Preservation Metadata, katso http://www.loc.gov/standards/premis/ ja http://www.oclc.org/research/projects/pmwg/), joka tätä nykyä on yleisesti hyväksytty alan standardi, on loistava saavutus. Standardin kehittäjät, päävastuullisena OCLC, eivät edes yrittäneet lähteä liikkeelle olemassa olevista palikoista, vaan kehittivät ensin pitkäaikaissäilytykseen soveltuvan tietomallin, ja loivat tarvittavat tietoelementit tältä pohjalta. Ilmestyessään vuonna 2005 standardi kaappasi Digital Preservation Award -palkinnon, eikä syyttä – harva asia helpottaa PAS-järjestelmien toteuttamista tulevaisuudessa yhtä paljon kuin PREMIS. Tosin standardia osaa arvostaa kunnolla vain, jos on perehtynyt ennen PREMISin julkaisemista vallinneeseen kaaokseen.

Lopuksi

Pitkäaikaissäilytyksen metadatan kehittäminen on osa paljon laajempaa prosessia, jossa kirjasto sopeuttavat järjestelmänsä ja prosessinsa digitaalisiin aineistoihin. Tämä prosessi on pitkä ja varmasti osin kivuliaskin, koska kyse on myös kirjastopoliittisista valinnoista: millaista aineistoa me haluamme tai meidän pitää tarjota asiakkaille, ja millä tavoin? Millaista yhteistyötä teemme tässä tarvittavien järjestelmien tuottamisessa, kansallisella ja kansainvälisellä tasolla? Olemme hiljalleen siirtymässä perinteisistä aineistoista digitaalisiin, ja tähän muutokseen liittyviä teknisiä, juridisia ja muita haasteita ratkotaan varmasti vielä pitkään.

Lopputulos lienee kuitenkin se, että valtaosa asiakkaista käyttää etupäässä sähköisiä aineistoja, ainakin tieteellisissä kirjastoissa. Uusi aineisto hankitaan tässä muodossa, ja vanhoista kokoelmista käytetymmät osat digitoidaan ennemmin tai myöhemmin. Mutta kokoelmien digitalisoituminen edellyttää sitä, että tuotantoprosessit ovat kunnossa ja niissä otetaan esimerkiksi pitkäaikaissäilytyksen vaatimukset huomioon.

Kansalliskirjastossa tämä muutostyö on aloitettu vuonna 2008, ja olemme päässeet jo hyvän matkaa eteenpäin. Valmista tuskin tulee vielä vuosiin, sillä ensin meillä pitää olla esimerkiksi PAS-sovellus. Kansalliskirjastossa on kuitenkin jo

nyt kohtuullinen ymmärrys tavoitteista ja keinoista niihin pääsemiseksi. Tässä on ollut paljon apua siitä tekniseen metadataan ja METS-standardiin liittyvästä osaamisesta, jota on viiden viimeisen vuoden aikaan kertynyt Mikkeliin, kansalliseen digitointikeskukseen.

Suomessa opetusministeriön vetämä KDK-hanke on erinomainen keino edistää digitaalisen kirjaston syntyä. Iso osa tarvittavasta kehitystyöstä ja sovellushankinnoista tehdään vuosina 2008–2013 KDK-sateenvarjon alla. Yhteinen projekti helpottaa myös koulutusurakkaa, joka on mittava. Ei vain kirjastossa, vaan kaikissa muistiorganisaatioissa ja kaikilla organisaation tasoilla on ymmärrettävä, mitä digitaalinen kirjasto on, millaisista palasista se rakentuu, ja mitä tämän kirjaston käyttöönotto edellyttää. 🕮

Tietoa kirjoittajasta:

Juha Hakala, kehittämisjohtaja
Kansalliskirjasto
email. juha.hakala@helsinki.fi