TTA, PAS ja julkishallinnon standardisointi

Juha Hakala
Kansalliskirjasto
2013-03-19

Miksi TTA tarvitsee KDK:n PAS-määrityksiä?

- Jos tutkimusdata on tarkoitus säilyttää pitkäaikaisesti (yhteisessä) PAS-sovelluksessa, datan on oltava KDK-määritysten mukaista
- Jos bittitason säilytys (ei migraatioita) riittää, KDK-vaatimuksia ei ole tarpeen / pakko noudattaa
- Haaste: pitkäaikaissäilytys pitäisi ottaa huomioon alusta lähtien, eli jo silloin kun tutkimusdata luodaan
  - Käytännössä tutkimusdatan pitkäaikaissäilytykseen kiinnitettäneen ”ruohonjuuritasolla” harvoin riittävästi huomiota, koska erityisosaamista ja resursseja asiasta huolehtimiseen on niukasti

KDK-määritykset (2013-03-18)

- Säilytys- ja siirtokelpoiset tiedostomuodot
  http://www.kdk.fi/images/kdk-pas-tiedostomuodot_v1.1.pdf (2013-01-22)
- Hallinnolliset ja rakenteelliset metatiedot ja aineiston paketointi
  http://www.kdk.fi/images/tiedostot/KDK_metatiedot_ja_aineiston_paketointi_v1.2.pdf (2013-02-28)
- Kansallisen digitaalisen kirjaston standardisalkku
  http://www.kdk.fi/images/KDK_standardisalkku_14.2.2013.pdf (2013-02-14)
- Varmista että käytössäsi on uusin versio!

KDK-määritysten ylläpidosta

- Vastuutaho KDK:n ohjausryhmä; käytännön työtä tehdään KDK:n teknisessä jaoksessa, erilaisissa työryhmissä sekä CSC:ssä, Kansalliskirjastossa ja muissa KDK-organisaatioissa virkatyönä
- Vastaavaa pitkäaikaissäilytykseen liittyvää ohjeistokokonaisuutta ei julkishallinnossa ole
- Sisältö rakentunut kirjastojen, arkistojen ja museoiden näköiseksi, uudet ”aluevaltaukset” kuten paikkatieto tai tutkimusdata edellyttävät laajennuksia

TTA ja KDK-tiedostomuotosuositus

- KDK:n määrittely kattaa tekstin, äänen, kuvan, elävän kuvan ja verkkoarkiston; muut aineistotyypit kuten tietokannat ”määritellään myöhemmin” tarpeen mukaan
- Määrittely erottaa toisistaan siirto- ja säilytyskelpoiset tiedostomuodot
  - Esimerkiksi Officen OOXML on siirto- mutta ei säilytyskelpoinen
- Määrittelyyn sisältyy myös ohjeistus kuvaan, ääneen ja tekstiin liitettävästä teknisestä metadatasta sekä tiedostomuodon pysyvyyden arviointikriteerit, joita soisi käytettävän julkishallinnossa laajasti

Tutkimusdatan haasteita

- Vaikka tutkimusdata olisi tekstiä, kuvaa tai ääntä, tekniset metatiedot voivat olla puutteelliset eikä tiedostomuoto välttämättä ole edes siirtokelpoinen
- Tietokantojen ja monien muiden aineistotyyppien siirtämiseen ei ole (yleisesti hyväksyttyä) tiedostomuotoa
- KDK:lle voi riittää ulkoasun säilyminen; tutkimusdatan vaatimustaso on usein kovempi
  - Excel-taulukon muuttaminen PDF:ksi voi olla turmioksi
- Vaikka data olisi periaatteessa yksinkertaista, sen tulkinta voi vaatia (pitkäaikaisssäilytettävän) ohjelmiston

TTA ja KDK:n paketointiohje

- KDK:ssa PAS-järjestelmään siirrettävä aineisto paketoidaan siten, että datan mukana lähetetään kaikki tarpeellinen metatieto METS-standardin mukaisesti
  - Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/
- KDK on laatinut METS-profiilin, joka määrittelee miten tätä monipuolista standardia tulee soveltaa
  - METS-pakettien tuottamista on harjoiteltu pilotissa; KDK-organisaatioillakin on vielä asiassa oppimista

Miksi datan paketointi on tarpeen?

- METS-paketti on säiliö (”kontti”), jonka avulla suurten aineistomäärien siirtäminen PAS-sovellukseen, muuttaminen säilytyskelpoiseen muotoon, migraatiot sekä jakelu PAS-sovelluksesta takaisin siirtäjäorganisaatioille ovat toteutettavissa tehokkaasti
- Ilman paketointia aineiston käsittelyä PAS-sovelluksessa olisi vaikeaa saada tehokkaaksi / automatisoida
  - Vrt. satamat ennen ja jälkeen konttiliikennettä
- METS-standardia sovelletaan lähes kaikissa kirjastojen PAS-hankkeissa ja yleisesti myös digitointiprojekteissa

Haasteita

- METS-paketteja on työlästä luoda käsin; tarvitaan ohjelmia jotka huolehtivat asiasta
  - METS-pakettien luonti pitäisi saada osaksi tutkimusdatan tuotantoprosessia (vrt. METS-pakettien luonti julkaisuarkistoihin tallennetuista aineistoista)
- KDK:n METS-profiilin mukaisen datan tuottaminen ei ole yksinkertaista, ainakaan aluksi
  - Jokaisessa METS-paketissa on oltava tietyt hallinnollisen metatiedon tietoelementit, joita ilman pitkäaikaissäilytys ei ole mahdollista
    - Esimerkiksi tiedostoista täytyy laskea tarkistussummat
- Pitkäaikaissäilytyksen tärkeyden ja säilytykseen liittyvien ongelmien ja niiden ratkaisutapojen ymmärtäminen

TTA ja KDK:n standardisalkku

- Standardisalkku määrittelee muun muassa metatietoformaatit, ohjelmistorajapinnat ja tunnistestandardit, joita KDK:ssa mukana olevien organisaatioiden sovellusten tulisi tukea
- Määrityksen taustalla ovat KDK:n yhteisten järjestelmien (asiakasliittymä ja PAS-sovellus) ylläpidon tarpeet sekä kirjastojen, arkistojen ja museoiden (digitaaliset) aineistot
- Tutkimusdataa ei ole vielä otettu huomioon, koska muistiorganisaatioilla ei ole sitä kokoelmissaan
  - Joillakin osa-alueilla tämä lienee helppoa, joillakin muilla ei

KDK-salkku ja tutkimusdata

- Kuvailevan metatiedon formaatti
  - TTA:n minimimetatietomalli perustuu Dublin Coreen, joten sen lisäämiseen salkkuun on matala kynnys
  - -> Lisäys luo perustan tutkimusdatan metatietojen viennille Finnaan
- Teknisen metatiedon formaatti
  - Ei tuottane ongelmia jos tutkimusdata on tekstiä, kuvaa tai ääntä; muut aineistotyypit kuten tietokannat ovat haasteellisia koska niille ei ole olemassa valmista teknisen metadatan formaattia
  - Pitäisikö KDK:n selvittää mahdollisuudet sellaisen kehittämiseen?
- Ohjelmistorajapinnat: OAI-PMH metatiedon haravointiin tutkimusdatan taustajärjestelmästä Finnaan

KDK-salkku ja tutkimusdata (2)

- Tunnukset ovat vaikeimmin sovellettava KDK-salkun osa-alue, koska niiden antaminen tutkimusdatalle on vaikeampaa kuin tunnusten jakelu esimerkiksi julkaisuille
  - Tarvitaan yleisiä linjauksia siitä mitä tunnuksia käytetään ja millä tavoin; sen jälkeen voidaan rakentaa tekninen toteutus (tunnusten jakelujärjestelmä)
- Tutkimusdatalle pitäisi soveltaa ns. pysyviä tunnisteita (Persistent identifiers), muun muassa koska Handle-, DOI- ja URN-tunnusten käyttöö suositellaan / vaaditaan alan kv. hankkeissa

Tutkimusdata ja PID-tunnukset

- Tutkimusdata on tunnistamisen kannalta hankalaa
  - Versioiden erottaminen toisistaan; tutkimusdata kuten tietokanta voi päivittyä jatkuvasti
  - Granulariteetti: pelkän tiedoston / datan tunnistaminen ei riitä, vaan usein pitää mennä hienojakoisemmalle tasolle (Clarin-hanke: käsitteet)
  - Standarditunnus puuttuu still-kuvilta, tietokannoilta,...
  - Tunnusjakelun vastuutaho puuttuu, kansainvälisesti ja kansallisesti
  - Tekijöiden (tutkijat / tutkijaryhmät / laitokset) tunnistaminen?
- Helpoimmin haluttuun tulokseen päästään oikoteitä, mutta tästä voi seurata harmia myöhemmin
  - Kun ratkaisuja tehdään, pitää ymmärtää mitä seurauksia niillä on; esimerkiksi oma tunnus voi muodostua ongelmaksi Finnassa

Esimerkki tunnuksien käytöstä

- Yliopiston laitoksella on itse laadittu MS Access –tietokanta, jossa on useita tuhansia eläimistä otettuja kudospreparaatteja kuvaavia viitteitä
  - Tietokantaan tallennetut kuvailut ymmärtää vain asiantuntija
- Joka viitteeseen liittyy kuvia, jotka on tallennettu tiedekunnan verkkopalvelimelle tutkijoiden omiin hakemistoihin
- PAS-luovutusta varten tarvittaisiin tunnus jokaiselle viitteelle ja siihen liittyville kuville; niitä olisi hyvä antaa myös esim. henkilöille (preparaatin tekijä, tuloksen tarkastaja etc.)
  - Yhden tietokannan saattamiseen PAS-kelpoiseksi voidaan tarvita kymmeniä tuhansia tunnuksia ja runsaasti teknistä metadataa

Tutkimusdata ja PID-tunnukset (2)

- Välttäkää epästandardeja ratkaisuja!
  - Esim. PID, jonka (oikea) tulkitseminen ei onnistu ilman sitä varten kehitettyä resolverisovellusta
- Valitettavasti kv. tutkimusdatahankkeissa on kehitetty toisinaan myös yleisen URI-syntaksin vastaisia ratkaisuja
- KDK:n linjausten mukaisesti Suomessa tulisi käyttää URN-tunnuksia; sen nimialueista tähän soveltuvat esimerkiksi URN:NBN, URN:OID ja URN:UUID.
  - URN:NBN suositeltavin ratkaisu
  - Tarvitaan lisäohjeistusta tutkimusdatan tunnistamisesta (ja laajemmin tutkimusdatan kuvailusta)

Toimijoiden (nimien) standarditunnus?

- Julkisiin avoimiin tietojärjestelmään soveltuvia vaihtoehtoja on kaksi, jotka molemmat ovat KDK-salkussa mukana
  - ORCID
  - International Standard Name Identifier, ISNI
- ORCID on kustantajavetoinen hajautettu järjestelmä
- ISNI on tekijänoikeusjärjestöjen ja kirjastojen täysin keskitetty järjestelmä jota sovelletaan takautuvasti
- Tunnusten rakenne (syntaksi) on sama, mutta muuten niiden välinen yhteistyökuvio on auki
  - Käyttöönotto Suomessa?

Juridisten henkilöiden tunnukset

- ORCIDia voivat hakea vain luonnolliset henkilöt; ISNI:n voi saada myös yliopisto, tutkimusryhmä, laitos tai muu (tieteellinen) julkaisijayhteisö
- Liiketoimintaa harjoittavat tahot saavat Y-tunnuksen, yhdistykset yhdistysrekisterinumeron
- Oleellista on kyetä jäljittämään vastuutaho
  - Kuka vastaa nyt tutkimusdatasta joka tuotettiin 50 vuotta sitten?
- Tunnuksen ohella tarvitaan ajantasaiset auktoriteettitiedot joissa on kuvattu esim. laitosten edeltäjät / seuraajat

Julkishallinnon standardisointi

- Taustalla tietohallintolakiin perustuva keskitetyn ohjauksen kehittäminen
- Koko julkishallintoa ohjaa ja tukee JHS-standardisalkku
  - https://www.yhteentoimivuus.fi/aihealue/Standardit
- Salkkua ylläpitävän JHS-standardisalkkuryhmän mandaatti perustuu JHS-suositukseen JHS 181
- JHS-salkkua täydennetään alakohtaisilla salkuilla (paikkatieto, KDK); niitä ylläpitävien ryhmien mandaatin perusta on suosituksen JHS 136 (Menettelytavat JHS-työssä) luvussa 6.1.3
  - Alakohtaisista salkuista voidaan nostaa standardeja JHS-salkkuun
- Valtiovarainministeriö voi päättää JHS-suosituksen muuttamisesta julkisen hallinnon tietohallinnon standardiksi

KDK-salkku ja JHS-standardisalkku

- Keskeisiä PAS-standardeja on jo nyt nostettu JHS-salkkuun (esimerkiksi METS ja URN-tunnus)
- KDK-salkku ja KDK:n tiedostomuotosuositus on tarkoitus lisätä JHS-salkun alisalkuksi seuraavassa JHS-standardityöryhmän kokouksessa
  - Suositukseen lisättäneen Maanmittauslaitoksen aloitteesta uusia, paikkatietoon liittyviä tiedostomuotoja
  - KDK-linjaukset voidaan ottaa käyttöön muilla hallinnonaloilla; valitettavasti KDK-organisaatioiden mahdollisuudet opastaa ja ohjeistaa julkishallintoa yleensä lienevät vähäiset; voimavarat menevät omien kaadereiden kouluttamiseen

Lopuksi

- Tutkimusdatan kuvailu minimimetatietomallin mukaisesti pitäisi olla varsin helppoa ja motivoivaa; porkkanana on aineiston löytyminen Finnasta
- Tutkimusdatan hallinnollisen metatiedon tuottaminen KDK-määritysten mukaisesti ei onnistu manuaalisesti, vaan edellyttää tuotantoympäristöä jossa PAS-vaatimukset on otettu huomioon - sekä valistuneita käyttäjiä
  - Dataa voidaan korjata PAS-kuntoon jälkikäteen vain rajoitetusti
- Jo tutkimusprojektia suunnitellessa pitäisi tietää, onko sen tuottamaa dataa tarpeen säilyttää pitkäaikaisesti, koska pitkäaikaissäilytys on ennakoitava tuotantoprosessissa