

Kokemuksia

*aineistojen paketoinnin piloteista.
KDK-pitkäaikaissäilytys 2013 -seminaari*



YHTEISKUNTATIETEELLINEN TIETOARKISTO
FINLANDS SAMHÄLLSVETENSKAPLIGA DATAARKIV
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Yhteiskuntatieteellinen tietoaarkisto

- > OKM:n rahoittama valtakunnallinen palveluresurssi
 - erillisyksikkö Tampereen yliopiston yhteydessä
 - perustettu jo 1999
 - toimikenttä laaja, ei vain yhteiskuntatieteet
- > Tehtävä: sähköisten tutkimusaineistojen pitkäaikaissäilytys ja toimittaminen jatkokäyttöön
 - kvantitatiivisia aineistoja ~1200, kvalitatiivisia ~150
 - 2012 toimitti lähes 700 aineistoa tutkijoille, opettajille ja opiskelijoille



Yhteiskuntatieteellinen tietoaarkisto

- > Toiminnan perustana asiantunteva, nopea, tasapuolinen ja maksuton palvelu
 - apua tutkimusaineistojen haussa ja käytössä
 - neuvoja aineistojen tietosuojakysymyksissä
 - ohjeita aineistonhallintaan ja pitkäaikaissäilytykseen
 - palvelee kansallisesti ja kansainvälisesti
 - verkostoitunut kansallisesti ja kansainvälisesti
 - mukana KDK:ssa 2008 alkaen, TTA:ssa 2012 →

Osana KDK:ta FSD

- > Eroaa “perinteisestä” arkistosta
 - aineiston laadun suhteen: tutkimuspainotus, datamatriisit...
 - käsittelyprosessin suhteen: aineistoja muokataan, originaali ei useinkaan “säilyttämisen arvoinen”
- > Silti aineisto poikkeaa tiedostomuodoiltaan sängen vähän KDK:ssä määritellyistä
 - .doc(x), .pdf, .rtf, .por, xml, .html, .jpg, .txt...
 - laadullinen aineisto sisällöltäänkin kulttuurista

Pilotti syksyllä 2012

> Alkuperäinen tavoite:

- tunnistaa tarpeelliset metadataelementit aineistojemme kuvauksista ja muodostaa manuaalisesti luovutuspaketti
- arvioida METS- yms. määntyksiä tutkimusaineistojen näkökulmasta

> Ja lopputulemana:

- muodostettiin luovutuspaketti ohjelmallisesti
- tutkimusaineiston näkökulma ei noussut keskeiseksi kysymykseksi pilotoinnissa

Pilotin eteneminen

- > Elementtien tunnistaminen oli melko helppoa, ja pakolliset elementit saatavilla metadatasta
 - silti perusteltua kerätä enemmän metadataa, esim. teknistä ja/tai aineiston käsittelyä kuvaavaa
 - “Millä ohjelmalla tai sen versiolla aineistoa on käsitelty?” “Kuka sen teki?!?”
 - Mitä tiedostopääte lopultakin kertoo tiedoston “laadusta”?
 - metadatan hallinta korostuu, mitä enemmän sitä on

Pilotin eteneminen

- > Ei pilotoitu käyttöoikeustietojen upottamista
 - ei saatavilla kattavaa ohjeistusta
- > Ei pilotoitu METS-paketin provenienssipuolta (säilytyssuunnitelmaa)
 - ei vielä valmista ohjeistusta → olisi keskeinen osa sekä teknisenä että sopimusluontoisena asiakirjana!
 - suunnitelma reflektoi osallistuvan organisaation tarpeita, aineistoa, metadataa ja asiakkuuksia

Pilotin eteneminen

- > Pilottiin panostettiin reilut kolme työviikkoa
 - päästy selvyyteen siitä, mitä METS on syönyt ja mitä sille pitää syöttää
 - päätös siitä, mitä tallennetaan ja mistä
 - paketoitiin 8 aineiston näyte
 - laadittu pilottia varten ohjelma, joka koosti luovutuspaketin joukosta esimerkkiaineistoja
- > Onnistuu tässä ajassa, kun käytettävissä XML-
asiantuntija... muuten jatkuisi vielä! 😊

Pilotin tuloksista

- > Tutkittiin laajasti ja kriittisesti saatavilla olevia määrittelydokumentteja
 - Ajatus: jos näistä on lähdettävä, niin virheitä tai ristiriitaisuuksia, tai sisäpiirin ymmärrystä vaativia osia ei saisi olla.
 - Havainto: asiakirjat vielä ”raporttimaisia”, eivät hyödyntäjälle helppoja. Myös asiakirjojen väliset suhteet vaativat tarkennusta. Terminologia vaatii tarkennuksia, ja määrittelyssä on vielä puutteita.
- näitä korjattu/korjataan uusiin versioihin

Poimintoja pilotin tuloksista

- > Vaihtoehtoisia esitystapoja pitäisi karsia
 - selkeyttää, kun on vain yksi tapa tehdä
- > Luovutuspaketin elinkaaren jokainen versio tulee olla uniikisti tunnistettavissa
 - tunniste + aikaleima
- > KDK-METS skeema ei vielä valmis
 - sen hyödyntäminen tuonee esille uusia haasteita, myös osallistujaorganisaatioissa

Luovutuspaketti

- > Koostettu paketti lähetettiin onnistuneesti
- > Mukana por- ja rtf-tiedostoja, jotka tiedettiin ei-sallituiksi → tietoinen valinta
- > Virhe terminologian tulkinnassa:
 - Seurauksena luovutuspaketti oli sijoitettu zip-paketin juureen, ei ensimmäisen tason juurihakemistoon
 - korostaa täsmällisyyden tarvetta spesifikaatiossa!

Johtopäätökset

- > Hyödyllinen seikkailu
 - perehtyminen METSiin ja PREMISiin bonusta
 - oman metadataformaattimme (DDI) peilaaminen näitä määrittämiä vasten, sekä sen validointi
 - periaatekeskustelu siitä, mikä on uniikkia pitkäaikaissäilytettävää, ja mikä kenties työekonomisista syistä säilytettävää, vaikka uudelleentuotettavissa
 - esimerkiksi alkuperäisaineiston käännös

Säilytettävän aineiston valinta

- > Pilotissa helppoa, mutta perustuttava tietoiseen valintaan
- > Tietoarkiston ohjaava asiakirja on arkistonmuodostussuunnitelma jota vasten, tai jota päivittäen, säilytyssuunnitelma tulee laatia
 - www.fsd.uta.fi/fi/hallinto/asiakirjat/AMS/ams_index.html

Takaisin johtopäätöksiin

> METADATAA

- Lisää! Parempaa! Hallittavassa muodossa!
Koneellisesti käsiteltävissä! Standardien mukaista!
Koko elinkaaren ajan!

> Palvelut, ohjelmistot ja apuvälineet luovutuspaketin muodostamiseen

- vain n. 70 Mt, alle 1000 tiedostoa: silti tuhansia rivejä
METS-pakettissa... Ei känsäisin käsin tuotettavissa.

> Rajapintojen merkitys korostuu

Kiitos!

- > Arkistoinnin, pitkäaikaissäilyttämisen ja uudelleen hyödyntämisen onnistuminen edellyttää aineiston hyvää hallintaa koko elinkaaren ajan.
- > *Tietoarkisto turvaa aineistosi sekä tukee tutkimusdatan avointa saatavuutta ja informoitua jatkokäyttöä.*

Tuomas J. Alaterä
tuomas.alatera@uta.fi