Avointen aineistojen julkaisualusta

Tanja Kantola, AVAA-projekti, CSC, 21.5.2013

AVAA-projekti

- OKM:n AVOIN-hankkeen osaprojekti CSC:llä 2/2013 – 3/2015
  - Vaatimusmäärittely-/ konseptointivaiheessa

- Toimeksianto: ”Toteutetaan avointen aineistojen **julkaisualusta**, joka **tukee** TTA:n kokonais-arkkitehtuurissa kuvattua tietoaineistojen **julkaisuprosessia**.”
  - Tietoa aineistojen julkaisun tukemiseen
  - Työkaluja prosessin automatisointiin
  - Liittymiä nyk. palvelujen välille

Tietoaineiston julkaisuprosessi

Tarkoitus	Prosessoitu, laatuvarmistettu ja metadatalla varustettu tietoaineisto julkaistaan käytettäväksi yleisesti tai rajoitetusti. Mahdollistetaan myös muualla (esim. kansainvälisessä tieteenalan omassa "pakollisessa" säilytyspaikassa) sijaitsevan datan käyttöönotto.
Sisältyvät osaprosessit	• Tietoaineiston luettelointi katalogiin • Tunnisteen antaminen tietoaineistolle
Lopputulos	Tietoaineisto, jonka voi saattaa uudelleen käyttöön. Data on julkisesti tai ainakin mahdollisimman laajan käyttäjäjoukon saatavilla selkeässä ja helppokäyttöisessä muodossa.

(Lähde: Tutkimuksen tietoaineistot – kokonaisarkkitehtuuri, v. 0.94)

Tarkennuksia

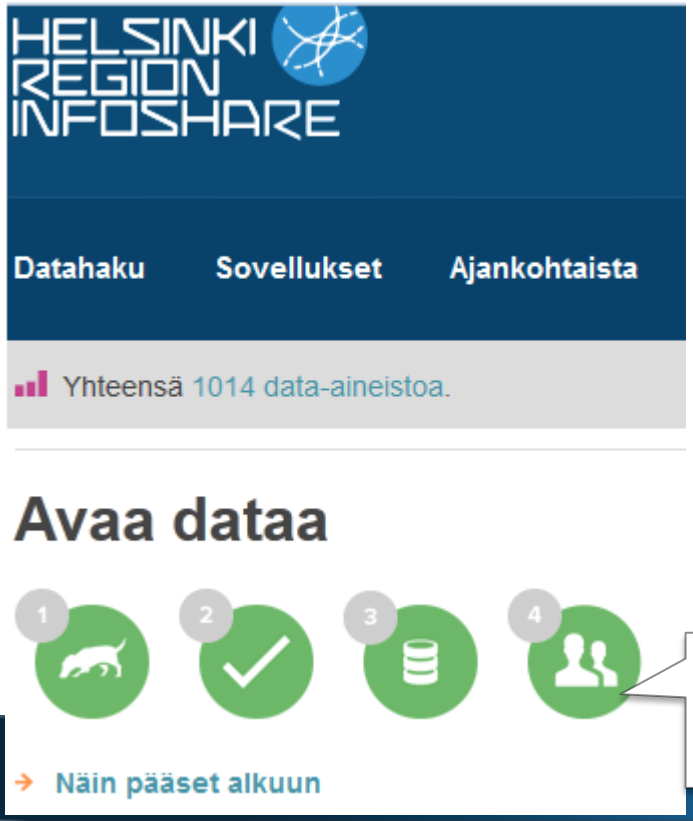
- ’Tietoaineiston julkaisu' → **datan** julkaisu
  - ei viittaa artikkeleihin tm. tutkimusjulkaisuihin, joista käytetään termiä ”julkaisu”

- AVAA ei ota kantaa aineiston tallennuspaikkaan
  - Tieteenalan omat (kans.väl. open access) säilytyspaikat
  - TTA:n tallennuspalvelut (IDA, PAS)
  - Yliopistojen tai kirjastojen (open access) repositoryt
  - …

Miten tukea tutkijoiden aineistojen avointa julkaisua?

Ohjeistus ja tiedonlevitys

- Joitain julkaisuprosessin vaiheita tuetaan parhaiten ohjeistuksella
- Ennen teknisiä ratkaisuja – ja niiden hyödyntämiseksi – tarvitaan valistusta (”infoportaali”).



- Tarjotaan helposti löydettävät, selkeät step-by-step-ohjeet ja tietopaketti parhaista käytännöistä: ”näin julkaiset tietoaineistosi”.
- Hyödynnetään muualla julkishallinnossa tehtyä työtä – suosituksia, ohjeita ja hyviä esimerkkejä aineistojen avaamiseksi
- Huomioidaan tutkijoiden aineistojen julkaisun erityistarpeet (mm. viittaukset julkaisujen ja aineiston välillä)

Esimerkki: selkeät ohjeet tukevat ja motivoivat aineistojen julkaisua. (http://www.hri.fi/fi)

Julkaisuprosessin vaiheita

1. Valitse, valmistele aineisto
2. Liitä aineistoon avoin lisenssi
3. Tallenna vakiintuneeseen (open access) tallennuspaikkaan
4. Luetteloi aineistokatalogiin, anna pysyväistunniste (PID)
5. Julkaisun jälkeen: hae, käytä, yhdistele aineistoja, viittaa niihin

Valitse ja valmistele aineisto

- Esim 1: CERN CMS: max 50% CMS-kokeen aineistosta julkaistaan n. 3 vuoden viiveellä (embargo-jakso)
  - Kompleksi hiukkasfysiikan aineisto dokumentoidaan hyvin
  - Julkaistavan alkuperäisen formaatin lisäksi aineisto muunnetaan yksinkertaisempaan opetuskäyttöön soveltuvaan formaatissa
- Esim 2: T. Tutkija julkaisee artikkeliin liittyvän aineiston
  - Tutkija tarkistaa aineiston laadun ja oikeellisuuden
  - Muuntaa sen avoimeen, koneluettavaan formaattiin

- Tarvitaan tietoa formaateista (CSV, XML-pohjaiset..), kriteerejä aineistojen valintaan, ohjeita laatutarkistusten tekemiseen ja aineistojen kuvailuun
  - AVAAn scope: kaikki tieteenalat. Tarkemmat ohjeet laaditaan parhaiten alakohtaisesti (esim. eri tieteenalojen XML-pohjaiset formaatit ja ohjelmistojen formaattimuunnostyökalut)

Liitä aineistoon lisenssi

- Lisenssi kertoo aineiston tuottajan ja käyttäjän oikeudet – millä ehdoilla aineisto on laillisesti käytettävissä
- AVAA: tietoa lisensseistä, mahdollistetaan koneluettavan lisenssin liittäminen aineistoon helposti (seurataan mm. JHS-työtä ”Avoimen datan lisenssimalli”), luvallisuus aineiston hakuehtona
- TTA suosittelee Creative Commons –lisenssiperhettä (”vahva ehdokas”, ei pakollinen)
- Panton Principles for Open Data in Science:
  - Valitse mahd. avoin lisenssi: kaupallista tai muuta uudelleenkäyttöä ei tulisi rajoittaa
  - Suositus erit. PDDL tai CCZero: ”Luovun kaikista yksin-oikeuksista lain sallimissa rajoissa” (suomennos meneillään)
  - Esim. avoin CMS-data julkaistaan CC0-lisenssillä

Luetteloi aineistokatalogiin

- … maailman löydettäväksi
- AVAA: liittymä TTA:n KATA-aineistokatalogiin
- Aineiston kuvailu metatiedoilla (viimeistään nyt) on välttämätöntä aineiston tulkinnan ja uudelleenkäytön mahdollistamiseksi
  - Aineiston alkuperä, muuttujien sisältö, käyttöehdot, laatu jne.
- TTA:n minimimetatietomalli: kaikille tutkimusaineistoille yhteiset vähimmäismetatiedot riippumatta aineistotyypistä, tieteenalasta tai tuotantotavasta.
- **Haaste**: Kaikki tutkijat eivät ole tottuneet datan kuvailuun ja kokevat sen vaikeana ja työläänä esim. minimimetatietojen auto-generoinnista huolimatta.
- Tarvitaan ohjeita ja esimerkkejä siitä, miten aineisto kuvataan tarpeeksi tarkkaan

Anna aineistolle PID

- Tutkimusaineistojen julkaisun erityispiirre
- Pysyväistunniste (PID), esim. URN tai DOI - kansainvälisesti uniikki tunniste säilyy samana vaikka aineiston sijainti vaihtuisi.
- Mahdollistaa viittaukset julkaisujen ja aineiston välillä.
  - Viittausmäärien kerääminen → meritoituminen
- TTA:n KATA ja IDA generoivat URN-tunnisteen tai käyttäjä voi antaa muualta hankitun tunnisteen lisätessään aineiston palveluun.

Julkaisun jälkeen…

- *Data on mahdollisimman avoimesti saatavilla selkeässä ja helposti uudelleen käytettävässä formaatissa*.
- Aineistojen tulisi löytyä helposti
  - Esim. Linked Open Data –menetelmien soveltaminen aineistoihin
- Aineistojen hyödyntämisen tulisi olla helppoa
  - AVAA: työkaluja aineistojen lataamiseen, rajapintajakeluun, tarkasteluun ja vakiomuotoisiin visualisointeihin
  - Pilottisovellukset, mm. avoimen CMS-datan hyödyntäminen lukio-opetuksessa
- Avoin julkaisu edistää tutkijan ansioitumista - työhön kertyy viittauksia, kun se on globaalisti saatavilla.

Haasteita

- Miten tästä tehdään tutkijoille helppoa ja vaivatonta & saadaan tutkijat toimimaan näin?
  - Pelkkä keppi ei riitä pitkään → tarvitaan toimintakulttuurin muutos – kollektiivinen oivallus aineistojen hyvän hallinnan, kuvailun ja avoimen julkaisun hyödyistä tutkijalle itselleen
- Tutkijat kokevat usein, ettei aineiston julkaiseminen ”maksa vaivaa”, mutta palkinto hyvin dokumentoidusta avoimesta aineistosta voi olla suuri
  - Maine ja työn vaikuttavuus, palaute, muiden motivointi datan jakamiseen, aineiston kumuloituminen ja parantaminen…
- Datan kuvailu vaatii ennemmin järjestelmällisemmän työskentelytavan omaksumista kuin ”ylimääräistä” vaivaa – merkataan kuvailutiedot jo aineiston syntyvaiheessa
  - Vanha dokumentoitu aineisto on arvokasta- hyöty esim. 5 vuoden kuluttua tutkijalle itselleen

Tietoa

- Open Knowledge Foundation: Open Data Handbook, http://opendatahandbook.org/pdf/OpenDataHandbook.pdf
- UNESCO: Ocean Data Publication Cookbook http://classroom.oceanteacher.org/pluginfile.php/6842/mod_resource/content/4/CookBook_MG64.pdf
- Panton Principles for Open Data in Science, http://pantonprinciples.org/
- Science Commons - Protocol for Implementing Open Access Data, http://sciencecommons.org/projects/publishing/open-access-data-protocol/
- Nine simple ways to make it easier to (re)use your data, https://peerj.com/preprints/7
- http://oad.simmons.edu/oadwiki/Data_repositories - a list of subject repositories for sharing data and long-term preservation
- ROAR – Registry of Open Access Repositories: http://roar.eprints.org/
- University of Oregon Library, Research Data Management Best Practices http://library.uoregon.edu/datamanagement/guidelines.html
- Australian National Data Services (ANDS), http://ands.org.au/services/index.html

Palautetta?

tanja.kantola@csc.fi