Metatieto – mihin ja miten?

Juha Hakala

Helsingin yliopiston kirjasto

juha.hakala@helsinki.fi

Sisältö

- Metatiedon määrittely
- Metatiedon käytöstä
- Metatietoformaatit
  - MARC, Dublin Core, IEEE LOM
- Elektronisten julkaisujen kuvailu ja säilytys
- Tallennus- ja hakuvälineistä

Metatieto

- Dokumentin rakenteinen kuvaus
- Voidaan tuottaa koneellisesti…
  - Google, Alta Vista, etc.
- ….tai käsityönä
  - kirjaston näyttöluettelo, kansallisarkiston arkistotietokanta, …
- Osa kuvailtavaa dokumenttia tai erillään
- Koskee julkaisua tai sen käyttöehtoja/ympäristöä

Metatiedon käytöstä

- Tiedonhaku, aineiston hankinta + evaluointi
  - suurimmissa kirjastojen tietokannoissa lähes 50 miljoonan teoksen tiedot; haku yhä tehokasta
  - LINDA = 3.7 miljoonaa tietuetta
  - Paikallistaminen (URL/paikanmerkki)
- Tekijänoikeuden metatiedot
- Pitkäaikaissäilytys (el. aineiston)

Metatiedon käytöstä (2)

- Tiedonhakuun tarvittavat elementit – kuten tekijä, nimeke, aihe jne.) tunnetaan, mutta
  - tekijänoikeuksien metadata on alkutekijöissään
  - pitkäaikaissäilytyksen kuvailutiedosta on monia ehdotuksia, mutta ei selkeää ratkaisua

Metatietoformaatit

- metatiedon rakenne perustuu formaattiin
- yksinkertaiset (Google), rakenteiset (Dublin Core) ja monimutkaiset (MARC) formaatit
  - vähäiset vs. kovat ammattitaitovaatimukset
- halvalla ei saa hyvää kuvailua, vaikka automaattinen indeksointi on kehittynyt
  - 1 MARC-tietue = 500 mk

Metatietoformaatit (2)

- formaatit yleensä domain-kohtaisia
  - kirjastojen MARC-formaattiin rakennettu sisään ”kirjastomainen” tapa käsitellä julkaisuja
  - Dublin Core pyrkii olemaan yleinen ratkaisu
- domaineilla omat kuvailusäännöt
  - monimutkaiset säännöt, monimutkainen formaatti (kehikko johon tiedot koodataan konelukuiseen muotoon)

MARC

- Machine Readable Cataloguing, kehitettiin Kongressin kirjastossa 70-luvun alussa
  - levinnyt kaikkialle maailmaan; kansallisia versioita (kuten FINMARC)
  - jokainen kirjastojärjestelmä MARC-pohjainen
- yliopistokirjastot siirtyvät FINMARCista kansainväliseen MARC21-formaattiin
  - tietojen kopiointi ulkomaisista kirjastoista

Dublin Core

- ANSI/NISO Z39.85-2001
  - amerikkalainen kansallinen standardi
- SFS-versio valmistuu syksyllä 2001
  - Suomessa jo versio asiakirjoille
- vastuutaho Dublin Core Metadata Initiative
  - http://www.dublincore.org/
- 14 työryhmää, joista yksi DC-Education
  - http://www.dublincore.org/groups/education/

Dublin Core (2)

- DC koostuu 15 kentästä ja joukosta (noin 70) niiden tarkenteita
  - http://www.dublincore.org/documents/dces/
  - http://www.dublincore.org/documents/dcmes-qualifiers/
- käyttöopas hyväksytty, suomennos –2001
- opas hyvin yleinen, hankkeiden kehitettävä omat ”luettelointisäännöt”
  - yhteismitallisuus kärsii tästä

Dublin Core (3)

- DC-tietojen tallennustapa määritelty HTML:, XML/RDF: ja XHTML:lle
- HYK ylläpitää tallennusalustaa
  - http://www.lib.helsinki.fi/cgi-bin/dc.pl
- haaste: tallennusvälineen ja tekstinkäsittelyohjelman yhdistäminen
  - erillistä tallennusalustaa vaikea käyttää

IEEE Learning Object Metadata

- Standardiluonnos, kehityksestä vastaa IEEE Learning Object Metadata Working Group
  - http://ltsc.ieee.org/wg12/
- taustalla IEEE Learning Technology Standards Committee (LTSC)
- monipuoliset tavoitteet
  - http://ltsc.ieee.org/wg12/s_p.html
- valmistumisajankohta avoinna

IEEE LOM & Dublin Core

- … ovat varsin yhteensopivia, katso
  - http://www.ischool.washington.edu/sasutton/IEEE1484.html ja standardiluonnoksen Annex B
  - …mutta IEEE LOM on paljon DC:tä laajempi
- tavoitteena yhteismitallisuuden parantaminen
  - http://dublincore.org/news/pr-20001206.shtml

Formaattikonversiot

- Periaatteessa mikä tahansa rakenteinen data voidaan muuntaa formaatista toiseen
- Käytännössä homma ei ole helppo
  - metatiedoissa olevat virheet, vaihtelevat nimien tallennustavat (etunimi sukunimi vai sukunimi, etunimi), kenttien yhteismitattomuus, merkkivalikoimien erot vaikeuttavat työtä

Elektronisen aineiston kuvailusta

- Painettuun aineistoon verrattuna:
  - aineistoa on hyvin paljon
  - dokumenttien määrittely ja rajaus vaikeaa
    - elektroninen lehti ”sulaa” verkkoon
    - elektroninen kirja voi koostua kymmenistä palasista
    - eri aineistot yhtenevät ”digimassaksi”
      - radiolähetyksessä kuvaa ja tekstiä
  - julkaisu voi paitsi lakata ilmestymästä, myös kadota täydellisesti

Elektronisen aineiston kuvailusta – ID-tunnukset

- Perinteiset ID-tunnusjärjestelmät kriisissä
  - ISSN ja verkkolehden määritelmä
  - ISBN:n kapasiteetin loppuminen
  - mikä tunnus oppimateriaaleille?
    - kansallisbibliografian ID-numero?
- Uudet tunnusjärjestelmät teoksille
  - ISAN: International Standard Audiovisual Number
  - ISWC: International Standard (Musical) Work Code
  - ISTC: International Standard Textual Work Code

Elektronisen aineiston kuvailusta – metatieto ja verkkokauppa

- E-bisnes ajaa kehitystä: jokainen kaupan oleva tuote pitää löytyä verkkokirjakaupasta
- metatieto todistaa että henkilö x on teoksen a tekijä
  - keskeinen merkitys tekijänoikeusasioiden selvittämisessä
  - metatietojen tuottajan oltava luotettava

Elektronisen aineiston kuvailu kirjastoissa

- Vain murto-osa aineistosta käsiteltävissä perinteisellä tyylillä (luettelointi)
  - verkkolehdet, opinnäytteet, e-kirjat
  - laaja kirjastojen välinen yhteistyö tarpeen
- Aihehakemistot (virtuaalikirjastot) nopeaan kuvailuun
- Kokoteksti-indeksointi
  - verkkoarkisto, 1800-luvun sanomalehdet

Elektronisen aineiston säilytys

- Useita ongelmia
  - tallennusväline muuttuu lukukelvottomaksi
    - aineisto kopioitava ”suojaan”
  - dokumentin käyttöön tarvittava ohjelma lakkaa toimimasta uusissa koneissa
    - TEKO (DOS-teksturi) ja Windows

Elektronisen aineiston säilytys (2)

- Kolme ratkaisumallia
  - säilytä vanhat koneet (ei onnistu)
  - konvertoi vanhat dokumentit
    - kaikkea aineistoa ei voi konvertoida - esim. ohjelmistoista ei ole lähdekoodia
    - konvertointi voi muuttaa julkaisua
  - alkuperäisen käyttöympäristön jäljittely
    - DOS-emulaattori Windows NT:ssä
    - alkuperäisten ohjelmien säilytys ja käytön opettelu
    - emulaattoreiden kasautuminen

Elektronisen aineiston säilytys (3)

- Käytännön projekteja vähän
  - Kansalliskirjastojen NEDLIB-hanke
    - CD ROM -levyjen käyttösovellus
    - verkkoaineiston haravaohjelmisto
    - ”emulation test bed”
    - pitkäaikaissäilytyksen edellyttämät metatiedot
  - eLibin CEDARS-projekti
- HYK kehittänyt suomalaisen version säilytyksen metatiedoista

Tallennusvälineistä

- Tallennusvälineet suunnattu usein ammattilaisille (välittäjille)
  - kirjastojärjestelmät (MARC, Dublin Core)
- Kuvailuvastuu osin siirtymässä dokumenttien tekijöille
  - asiakirjojen hallintajärjestelmät, computer aided radio – järjestelmät
- Hyvät apuvälineet tarpeen
  - Vesa-verkkosanasto (http://vesa.lib.helsinki.fi)

Hakuvälineistä

- Verkon indeksointisovellukset kehittyneet nopeasti (Google, Alltheweb)
  - skaalautuminen verkon mukana
- Tulosjoukot liian suuria
- Metatietoa verkossa paljon, laatu ongelma
  - hakukoneiden rakentajat eivät ole panostaneet tehokkaaseen metadatahakuun
  - panostettava hyvän metatiedon tallennukseen

Virtuaaliyliopiston kirjasto?

- Tarvitaan sekä periaatepäätöksiä sekä käytännön työvälineiden kehittämistä
  - minkä aineiston kansalliskirjasto luetteloi ja tallentaa osana vapaakappalevastuutaan?
  - mistä materiaalista vastaavat yliopistot, ja miten esimerkiksi metatiedon luonti tehdään?
    - ”itsepalvelu” ainoa toimiva ratkaisu jos dokumentteja on todella paljon