

Dublin Core ja sen käyttö WWW-sivuilla

ITKD50

Tekstitiedonhaku

Workshop-seminaari

22.04.2004

Sami Äyrämö, Timo Aittokoski, Miika Nurminen

Sisältö

Metadata

Dublin Core –elementit

Dublin Coren käyttö www-sivuilla

Metadata I

- ”Tietoa tiedosta” (*engl. data about data*)
- Käyttötarkoituksia:
 - tiedon kuvaileminen ja identifiointi
 - käyttäytymisen, toiminnan ja käytön dokumentointi
 - dokumentointi tiedon suhteista muihin tietokokonaisuuksiin
 - dokumentointi siitä kuinka tietoa tulisi käsitellä
- Käyttäjiä mm.:
 - Tietojärjestelmien suunnittelijat, tekijät, ylläpitäjät, käyttäjät,...

Metadata II

- Hieman toisistaan eroavia merkityksiä eri käyttäjäkuntien keskuudessa
- Tavoite sisältöjen saatavuuden tehostaminen:
 - aineistojen järjesteleminen, luettelointi, kuvaileminen,...
 - esim. voidaan helpottaa Internetissä sivuston löytymistä
- Nykyään tärkeä rooli erityisesti kehitettäessä tehokkaita, skaalautuvia, asiantuntevia ja kestäviä tietojärjestelmiä
- Perinteisiä sovellusalueita esim. museorekisterit, kirjastoluettelot, arkistot,...

Metadata III

- *”Sum total of what one can say about any information object at any level of aggregation”*
- **Informaatio-objekti:**
 - Asiakokonaisuus johon voidaan viitata tai jota voidaan käsitellä erillisenä kokonaisuutena ihmisen/laitteen toimesta
- **Informaatio-objektiin liittyy:**
 - Sisältö: mitä objekti on, mistä se koostuu.
 - Konteksti: ulkoisia tekijöitä kuten: kuka, mikä, miksi, missä, kuinka objekti luotiin
 - Rakenne: objektien sisäisiä tai välisiä assosiaatioita

Metadataluokat

- **Hallinnollinen**
 - esim. sijaintitieto
- **Kuvaileva**
 - esim. hakupalvelut
- **Tekninen**
 - esim. laitteisto- ja ohjelmistodokumentointi
- **Metadatan säilyttäminen**
 - esim. tietojen päivitys ja muutokset
- **Metadatan käyttö**
 - esim. käytön ja käyttäjän jäljitys

Dublin Coren elementit

- elementtejä käytetään verkossa saatavilla olevien resurssien (yleensä dokumenttien) kuvailemiseen ja identifioimiseen
- elementtejä on 15 kappaletta
- elementit koostuvat nimi-arvo pareista, esim.
Title="The Sound of Music"

Elementtien käytöstä

- Jokainen elementti on vapaaehtoinen ja toistettavissa
- Elementtien metadata on suositeltavaa luoda tiettyjen suositusten mukaan
- Elementin arvo voidaan (ja kannattaa) tietyissä elementeissä valita "kontrolloiduista sanastoista"

Dublin Coren tarkenteet

- DC tasapainoilee kahden asian kanssa: DC on pidettävä yksinkertaisena mutta sen oltava myös tehokas ja tarkka apuväline tiedonhaussa
- Tarkenteilla elementtejä voidaan laajentaa, jotta se palvelisi paremmin vaativia tiedontarpeita
- Tarkenteet jaetaan kenttätarkenteisiin ja merkintäjärjestelmiin

Dublin Coren tarkenteet, 2

- Kenttätarckenne (element qualifier) on tarckenne, joka muokkaa kentän merkitystä yksityiskohtaisemmaksi
- Merkintäjärjestelmä (encoding scheme) on tarckenne, jolla viitataan jonkin standardin tai konvention mukaiseen merkintätapaan
- 15 peruskentän lisäksi formaatin soveltaja voi ottaa käyttöön omia kenttiä ja tarkenteita

Elementtien jaottelu

- Elementit voidaan karkeasti luokitella kolmeen eri ryhmään siten, että osa elementeistä liittyy kohteen sisältöön, osa tekijätietoihin ja osa kohteen tunnistamiseen (versioon)

Elementtien jaottelu, 2

Sisältöön liittyvät	Tekijätietoihin liittyvät	Versioon liittyvät
Coverage (Kate) Description (Kuvaus) Type (Laji) Relation (Suhde) Source (Lähde) Subject (Aihe) Title (Nimike) Audience (Kohderyhmä)	Contributor (Muu tekijä) Creator (Tekijä) Publisher (Julkaisija) Rights (Tekijänoikeudet)	Date (Päivämäärä) Format (Formaatti) Identifier (Identifikaatiotunnus) Language (Kieli)

Elementit taulukkona

<p><u>Title</u> (Nimeke)</p> <p>Tekijän, laatijan tai julkaisijan antama asiakirjan nimi, otsikko tai asia.</p>	<p><u>Alternative</u></p>
<p><u>Creator</u> (Tekijä)</p> <p>Henkilö tai yhteisö, joka on ensisijaisesti vastuussa tallenteen intellektuaalisesta sisällöstä. Esimerkiksi kirjoittaja (jos kyseessä kirjoitettu dokumentti), taiteilija, valokuvaaja tai kuvittaja (kyseessä kuvadokumentti).</p>	<p>-</p>
<p><u>Subject</u> (Aihe)</p> <p>Tallenteen aihealueen kuvaus, yleensä asiasanoilla tai fraaseilla kuvattuna. Kontrolloitujen sanastojen ja luokitusten käyttö on suositeltavaa.</p>	<p>-</p>
<p><u>Description</u> (Kuvaus)</p> <p>Tallenteen sisällön kuvaus vapaatekstinä. Esimerkiksi tekstidokumentin abstrakti (tiivistelmä) tai visuaalisen tallenteen sisällönkuvaus.</p>	<p><u>Table Of Contents</u></p> <p><u>Abstract</u></p>

Elementit taulukkona, 2

<p><u>Publisher</u> (Julkaisija)</p> <p>Organisaatio, joka on julkaissut tai julkistanut tallenteen sen nykyisessä muodossaan. Esimerkiksi kustannusyhtiö, yliopiston laitos tai jokin muu yhteisö.</p>	
<p><u>Contributor</u> (Muu tekijä)</p> <p>Henkilö tai organisaatio, joka Tekijä-kentässä mainittujen henkilöiden/organisaatioiden lisäksi on osallistunut merkittävästi tallenteen luomiseen, mutta jonka panos on kuitenkin ollut toissijainen verrattuna ns. päävastuullisen tekijän osuuteen. Esimerkiksi toimittaja, kääntäjä tai kuvittaja.</p>	
<p><u>Date</u> (Päivämäärä)</p> <p>Päivämäärä, jolloin tallenne on julkaistu tai julkistettu kyseisenä versiona. Tätä päivämäärää ei pidä sekoittaa Kate-kentässä esiintyviin aikamerkintöihin.</p>	<p><u>Created</u> <u>Valid</u> <u>Available</u> <u>Issued</u> <u>Modified</u> <u>Date Copyrighted</u> <u>Date Submitted</u></p>

Elementit taulukkona, 3

<p><u>Type</u> (Laji)</p> <p>Tallenteen laji, esimerkiksi kotisivu, romaani, runo, artikkeli, sanakirja tai valokuva. Kuvailun yhteismitallisuuden takia laji tulisi valita Dublin Core -yhteisön ylläpitämästä listasta, joka ei ole yhteismitallinen FINMARC-formaatin kirjallisuuslaji- ja teoksen sisältö-koodien kanssa.</p>	-
<p><u>Format</u> (Formaatti)</p> <p>Tallenteen tiedostoformaatti, kuten teksti/html, sovellus tai JPEG-kuva. Formaattitietoa voidaan käyttää sen selvittämiseen, mitä laitteisto- ja ohjelmistoympäristöä tallenteen hyödyntäminen edellyttää.</p>	<p><u>Extent</u></p> <p><u>Medium</u></p>
<p><u>Identifier</u> (Identifikaatiotunnus)</p> <p>Tunnus (merkkijono tai numero), joka yksiselitteisesti identifioi tallenteen. Esimerkiksi verkkodokumenttien URN tai URL-tunnukset (URL:n eli sijaintitiedon käyttöä ainoana identifikaatiotunnuksena ei suositella.)</p>	<p><u>Bibliographic</u></p> <p><u>Citation</u></p>

Elementit taulukkona, 4

<p><u>Source</u> (Lähde)</p> <p>Tunnus (merkkijono tai numero), joka yksiselitteisesti identifioi tallenteen lähteen, jos sellainen on olemassa.</p>	
<p><u>Language</u> (Kieli)</p> <p>Tallenteen kieli. Kieli tallennetaan ISO. 639-1 standardin mukaisena kaksikirjaimisena koodina. (Esim. en, fi, fr)</p>	
<p><u>Relation</u> (Suhde)</p> <p>Ilmaistaan toisen tallenteen identifikaatiotunnus, joka on jossakin suhteessa kuvailtavaan tallenteeseen. Elementissä voidaan myös ilmaista kuvailtavan tallenteen suhde muihin tallenteisiin (ks. suhdetyypit alla). Esimerkiksi dokumentin sisältämät kuvat, kirjan luvut tai kokoelman osat.</p>	<p><u>Is Version Of</u> <u>Has Version</u> <u>Is Replaced By</u> <u>Replaces</u> <u>Is Required By</u> <u>Requires</u> <u>Is Part Of</u> <u>Has Part</u> <u>Is Referenced By</u> <u>References</u> <u>Is Format Of</u> <u>Has Format</u> <u>Conforms To</u></p>

Elementit taulukkona, 5

<p><u>Coverage</u> (Kate)</p> <p>Tallenteen ajallinen tai maantieteellinen kate. Maantieteellinen kate viittaa fyysiseen alueeseen; koordinaatteihin tai paikannimeen. Ajallinen kate viittaa lähinnä tallenteen asiasisältöön, ei siihen koska se on luotu tai asetettu saataville (kuuluvat Päivämäärä-elementtiin).</p>	<p><u>Spatial</u></p> <p><u>Temporal</u></p>
<p><u>Rights</u> (Tekijänoikeudet)</p> <p>Lyhyt maininta tallenteen käyttöoikeuksista tai linkki tallenteen tekijänoikeustietoihin tai palveluun, joka antaa tietoja tallenteen käytöstä ja oikeuksista.</p>	<p><u>Access Rights</u></p>
<p><u>Audience</u> (Kohderyhmä)</p> <p>(Huom.! Tämä ei ole osa Simple Dublin Coren viidestätoista elementistä, ja tätä pitäisi käyttää vain jos käytössä on Qualified Dublin Core.)</p> <p>Ryhmä jolle tallenne on suunnattu tai katsottu hyödylliseksi. Kohderyhmän määrittää Tekijä, Julkaisija tai muu kolmas osapuoli.</p>	<p><u>Mediator</u></p> <p><u>Education Level</u></p>

Dublin coren käyttö WWW-sivuilla

- Esitystavat: HTML:n META- ja –LINK –elementit, XML:n nimiavaruus tai RDF-kuvaukset.
- Esitystavasta riippuen metatieto voi olla dokumentissa tai sen ulkopuolella.
- Tekstimuotoista, voidaan editoida käsin
- Suositeltavampaa käyttää jotain ohjelmaa, esim. HYK:n Dublin Core –tallennusalausta
 - <http://www.lib.helsinki.fi/cgi-bin/dc.pl>
- DMCI:n sivuilla lisää työkaluja
 - <http://dublincore.org/tools/>

META-elementit

- META-elementillä merkitään
(ominaisuus, arvo) –parit.
- LINK-elementillä viitataan dokumentin ulkopuolella oleviin resursseihin.
- META-elementtien väärinkäytön takia jotkut hakukoneet jättävät nämä huomiotta.
- Metatieto on sijoitettava kuvattavaan HTML-tiedostoon.

META-esimerkki

```
<head profile="http://dublincore.org/documents/dcq-html/">
  <title>Expressing Dublin Core in (X)HTML ...</title>
  <link rel="schema.DC"
        href="http://purl.org/dc/elements/1.1/" />
  <link rel="schema.DCTERMS"
        href="http://purl.org/dc/terms/" />
  <meta name="DC.title" lang="en"
        content="Expressing Dublin Core in (X)HTML ..." />
  <meta name="DC.identifier" scheme="DCTERMS.URI"
        content="http://dublincore.org/documents/dcq-html/" />
  <meta name="DC.format" scheme="DCTERMS.IMT"
        content="text/html" />
</head>
```

Profilessa määritetään, miten Dublin Coren elementit esitetään HTML:na. Schema-linkeillä kuvataan, elementtien ja tarkentimien nimiavaruudet.

XML-nimiavaruudet

- XML-dokumentti voi koostua useampaan skeemaan kuuluvista elementeistä. Skeemat erotetaan etuliitteillä.
- Dublin Core –määrittelykset voidaan kuvata omassa nimiavaruudessaan.
- Sovelluksen täytyy päätellä, mitkä elementit käsitellään metatietona.
- Metatieto on sijoitettava kuvattavaan XML-tiedostoon.

XML-esimerkki

```
<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/
                     http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>UKOLN</dc:title>
  <dc:description>
    UKOLN is a national focus of expertise in digital
    information management...
  </dc:description>
  <dc:identifier xsi:type="dcterms:URI">
    http://www.ukoln.ac.uk/
  </dc:identifier>
</metadata>
```

**Tarkentimet voidaan kuvata XML-skeeman XSI:type-attribuutilla
Metadata tekstinä, ei attribuutteina!**

RDF (Resource Description Framework)

- W3C:n kehittämä yleiskäyttöinen tietomalli resurssien kuvaamiseen (resurssi == mikä tahansa URI:lla viitattava olio, esim. WWW-sivu, henkilö, verkossa oleva laite). Yleensä XML-kielinen esitys.
- RDF-kuvaukset ovat kolmikkoja (resurssi, ominaisuus, arvo), joista jokainen voi myös olla resurssi.
- RDF-kieli tukee tietorakenteita ja väitteiden ilmaisemista toisista väitteistä (reification).
- Semanttisen webin perustekniikka.
- Metakuvaus resurssissa tai (yleensä) sen ulkopuolella.
- Verkkomainen tietorakenne (vs. XML:n puu).

RDF-esimerkki

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description
    rdf:about="http://www.cc.jyu.fi/~minurmin/luk/luk.pdf">
  <dc:title>Semanttiset verkot ...</dc:title>
    <dc:creator
      rdf:resource="http://www.cc.jyu.fi/~minurmin/" />
    <dc:subject><rdf:Bag>
      <rdf:li>käsitegraafit</rdf:li>
      <rdf:li>semanttinen web</rdf:li>
    </rdf:Bag></dc:subject>
    <dc:language>fi</dc:language>
  </rdf:Description>
</rdf:RDF>
```

Kuvauksissa voi olla rakennetta (esim. joukkojen kuvaus bag-tietorakenteella, linkit muihin resursseihin).

RDF-graafi

