



**Ontologiat ja semanttinen web sisällön  
tuotannon näkökulmasta  
Luetteloinnin tiedotuspäivä 2010**

**Juha Hakala  
Kansalliskirjasto**

KANSALLISKIRJASTO



NATIONALBIBLIOTEKET

**Sisältö**

- **Semanttinen Web ja kirjastot –yleistä**
  - W3C Libraries and Semantic Web Incubator Group
- **Semanttinen Web käytännössä: resurssien tunnistaminen ja ontologiat**
  - JHS Sanastot -työryhmä
  - Vocabulary Mapping Framework

KANSALLISKIRJASTO



NATIONALBIBLIOTEKET

## Semanttisen Webin visio

### •Tim Berners-Lee:

- “I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers” (1999).
- “Semantic web could bring about a revolution in how, for example, scientific content is managed throughout its life cycle” (2006).
- Kirjastojen kannalta semanttinen Web on merkittävä vaikka se toteutuisi vain osittainkin
- Kattava toteutus voisi olla ongelma:
  - Please-rob-me –palvelu; henkilövaaka ja jääkaappi

## Semanttisen Webin käytäntö

- Visiot ovat muuttuneet käytännöksi hitaasti
- ”SW-sateenvarjon” alla voidaan tehdä järkeviä ja käytännöllisiä asioita, kunhan tavoitteet asetetaan realistisesti
- W3C Libraries and Semantic Web Incubator group
  - The goal is to help lowering the barriers of Semantic Web adoption in the library community
- Esimerkkejä sovelluskohteista:
  - Resurssien tunnistaminen
  - Ontologioiden ja rakenteisen metadatan soveltaminen
- Kirjastojen kannalta asioissa sinänsä ei ole mitään mullistavaa, mutta asioiden toteuttamistapa ja toiminnan tavoitteet ovat uudenlaisia

## Resurssien tunnistaminen: linked data

### •Tim Berner-Lee (taas):

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that [the users] can discover more things.

## URLen käytöstä: cool URI vs PID

### •Kaksi koulukuntaa:

- Resurssien tunnistamiseen voidaan käyttää URL:ää, kunhan huolehditaan nimipalvelun avulla siitä että se ei koskaan muutu

- Aurinko-termin cool URI on <http://www.yso.fi/onto/yso/Y100928>

- Resurssien identifiointiin on paras käyttää PID-tunnistetta (URN, DOI, etc), joka linkitetään dokumentin URL-osoitteeseen / -osoitteisiin

- Konttinen, Pekka: Living on voles – plastic life of the Ural owl – julkaisun URN on <http://urn.fi/URN:ISBN:978-952-10-6114-1>

## Cool URI & PID: vertailua

- Cool URI-ratkaisu on teknisesti yksinkertainen, koska se ei edellytä tunnisteiden linkkaamista URL-osoitteeseen ns. resoluutiopalvelussa
- PID-ratkaisu on toiminnallisesti parempi, koska
  - yksi PID voidaan linkittää useisiin URL-osoitteisiin jos dokumentista on useita kopioita
  - resoluutiopalvelulta voidaan pyytää erilaisia asioita (dokumentti; sitä koskevat metatiedot; luettelo osoitteista joista dokumentti löytyy)
- PID-ratkaisu on luotettavampi, koska
  - domain-nimet voidaan myydä
  - Cool URI:a ei voi erottaa tavallisesta URL:stä

## Nykytilanne

- Kirja-ala ja kirjastot käyttävät PID-tunnisteita, suosituimmat vaihtoehdot ovat DOI ja URN
  - Mahdollisuus valvoa tunnisteiden käyttöä
  - PersID-hanke
- W3C ja useimmat semanttisen Webin hankkeet nojaavat cool URI –tunnisteisiin
  - Tiedon lisäksi myös tunnisteiden jakaminen vapaata
- Berners-Leen tavoitteet (use URIs as names of things; use HTTP URIs so that people can look up those names) toteutuvat molemmilla tavoilla

## Ontologiat

- Merkittävä rooli semanttisen Webin toteuttamisessa, niin Suomessa kuin muualla
- Laajapohjainen lähestymistapa a'la FinnONTO on toimivin ratkaisu, koska tarvitaan muun muassa:
  - Toimijaontologia (eli nimiauktoriteetit)
  - Paikkaontologia (maantieteelliset nimet)
  - Yleinen (suomalainen) ontologia ja sitä täydentäviä erikoisalojen ontologioita
- Lisäksi tarvitaan näiden ontologioiden jatkuvaa ylläpitoa ja sille tekninen alusta

## Ontologioiden soveltamisesta

- YSAn ja siihen pohjautuvien erikoisalojen sanastojen vuosikymmeniä jatkunut käyttö antavat hyvän pohjan suomalaiselle semanttiselle Webille, mutta sanastojen/ontologioiden käyttöä on kyettävä laajentamaan koko julkishallintoon
  - VM:n valmisteilla oleva "FinnONTO-laki"
- Kirjastot tarvitsevat erillisen nimiauktoriteettitietokannan, mutta sen lisäksi tarvitaan vielä julkishallinnon yhteinen järjestelmä

## JHS 170 & JHS 180(?)

- JUHTA (Julkisen hallinnon tietohallinnon neuvottelukunta) on valmstellut suositukset
  - Julkishallinnossa tiedonsiirtoon käytettävistä XML-skeemoista (JHS 170) sekä
  - Julkisen hallinnon sanastotyöprosessista (vielä julkaisematon teksti, luultavasti JHS 180)
- JHS 180 kuvaa prosessin jolla julkisen hallinnon toimijoiden tarvitsemat käsitteet kuvataan
  - Tavoitteena on julkisen hallinnon yhteisen metadatarokisterin luonti

## Julkisen hallinnon sanastotyöprosessista

- Tavoitteena ei ole perinteisten sanastojen ylläpitoprosessin vaan metadataelementtien määrittelyprosessin kuvaaminen
  - ”Miten toimitaan silloin, kun halutaan määrittellä ne tietoelementit, joita tarvitaan henkilötietojen siirtämiseen julkisen hallinnon toimijoiden välillä?”
    - Perustetaan työryhmä, joka...
- Siirrettävän tiedon syntaksi määrittellään JHS 170:ssä, tietosisältöjen määrittelyprosessi JHS 180:ssa
- Tavoitteena on vähentää julkisessa hallinnossa tehtävää päällekkäistyötä esimerkiksi henkilöiden tietojen tallennuksessa ja parantaa tietojen ”koneymärrettävyyttä”

## Metadatan ontologisointi

- Yksi Semanttisen webin haasteista on, että metadataa on tallennettu hyvin monissa eri muodoissa
  - Kansallinen digitaalinen kirjasto –hankkeessa käytetään 12 eri vaihtomaattia; miten saamme kaiken tämän metatiedon yhtenäistetyksi siten, että se saadaan tehokkaasti haettavaksi yhteen järjestelmään?
- Yksi vaihtoehto on Vocabulary Mapping Framework- eli VMF-hankkeen tulosten soveltaminen

## VMF:n lähtökohdat

1. *Bibliographic and heritage metadata is becoming increasingly diverse and complex and will require increasing interoperability for re-use and discovery*
  2. *Members of the JISC community have increasingly diverse, complex and unpredictable metadata requirements*
  3. *Metadata from producers/providers/publishers will become increasingly important as a substantial component of metadata in the JISC community*
- Kaikki nämä ovat relevantteja myös Suomessa

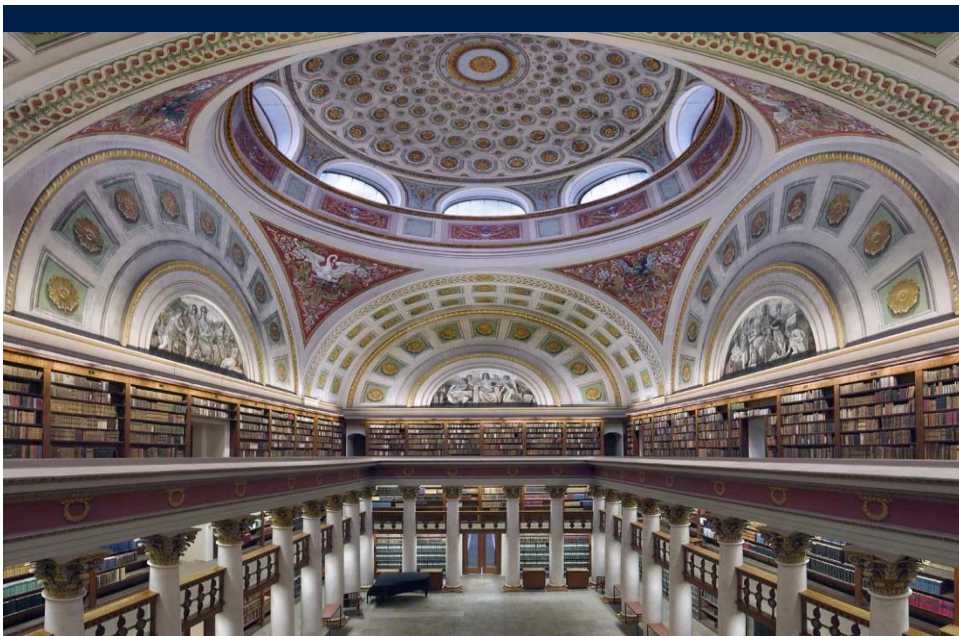
## VMF matrix

- VMF matrix
  - a tool which can be used to automatically compute the "best fit" mappings between terms in controlled vocabularies in difference metadata schemes
- Matriisi on ontologia johon sisältyy toistatuhatta termiä; on olemassa myös versio jossa termeihin on linkitetty muutamien formaattien (esim. MARC21, ONIX, DC) elementtejä, mikä mahdollistaa mappauksen teon
- Ontologia tultaneen asentamaan ONKI-palveluun, jossa siihen on mahdollista linkittää lisää metadataformaatteja; käyttö Suomessa on vielä auki

KANSALLISKIRJASTO



NATIONALBIBLIOTEKET



KANSALLISKIRJASTO



NATIONALBIBLIOTEKET