Kokoelmatietojen siirto MuseoSuomi-järjestelmään

Mirva Salminen
mirva.salminen@cs.helsinki.fi
Semantic Computing Research Group (seco)
University of Helsinki & Helsinki Institute for Information Technology (HIIT)

Esityksen sisältö

1. Sisällöntuotannon ongelma: lähtökohdat ja tavoitteet
2. Sisällöntuotannon perusratkaisu
3. Sisällön syntaktinen yhteensopivuus
4. Sisällön semanttinen yhteensopivuus
5. Annotointiprosessin kulku
6. Yhteeveto: käytäntö ja ongelmat

Sisällöntuotannon lähtökohdat

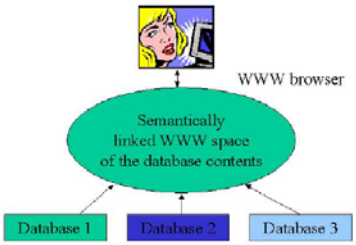
Kokoelmatiedot ovat heterogeenisiä:

1. Rakenteeltaan
   - Eri museoiden tietokannat ovat erillisiä
   - Eri museot käyttävät erilaisia tietokannan hallintajärjestelmiä
     - MS SQLServer, Oracle, jne.
   - Tietokantojen loogiset rakenteet – taulut, kentät, tietotyypit, jne. - eroavat
2. Sisällöltään
   - Erilaiset kokoelmat sisältävät erilaista tietoa
     - Esim. Muinaismuistoista on erilaista tietoa kuin huonekaluista tai valokuvista
   - Eri kokoelmia kuvaillaan eri tavoin ja eri sanastoilla
     - Esim. Toisaalla käytetään lelu-sanaa ja toisaalla leikkikalu-sanaa
     - Esim. Toisaalla krikettipelistä kerrotaan tarkalleen jopa pallojen, porttien ja mailojen lukumäärät
     - Esim. Ajan 1995-1999 voi kuvata myös määreellä “1990-loppupuoli” tai “1990 lp”

Sisällöntuotannon tavoitteet

- MuseoSuomi tarjoaa:
  - Yhtenäisen julkaisukanavan webiin erilaisille kokoelmille – museoesineille, taideteoksille, valokuville, muinaismuistoille
  - Sisältöperustaisen semanttisen tiedonhaun
  - Semanttisen suosittelun
- Sisällöntuotannon tavoitteena on:
  - Tuottaa kokoelmatiedoista sellainen paketti, joka mahdollistaa MuseoSuomen tarjoaman toiminnallisuuden

Sisällöntuotannon tavoitteet (2)



Esityksen sisältö

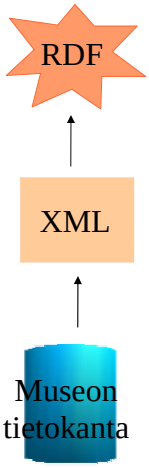
1. Sisällöntuotannon ongelma: lähtökohdat ja tavoitteet
2. **Sisällöntuotannon perusratkaisu**
3. Sisällön syntaktinen yhteensopivuus
4. Sisällön semanttinen yhteensopivuus
5. Annotointiprosessin kulku
6. Yhteenveto: käytäntö ja ongelmat

Ratkaisu: Museon tietokannasta MuseoSuomeen

RDF
- Resource Description Framework
- MuseoSuomen käyttämä tietomuoto

XML
- eXtensible Markup Language
- Valinta, mitä tietoja otetaan mukaan
- Tietojen esitys samanlaisessa muodossa

Museon tietokanta
- Kaikki tiedot
- Erilaiset tietokannat
- Erilainen terminologia



Semanttinen yhdistäminen
- Eri terminologioiden yhdistäminen Esim. Pikkupöydät ovat pöytiä ja lenkkarit ovat lenkkitossuja
- Resurssien tunnistaminen

Syntaktinen yhdistäminen
- Erilaisista tietokannoista samanlaiseen muotoon
- Esitettävien tietojen valinta

Esityksen sisältö

1. Sisällöntuotannon ongelma: lähtökohdat ja tavoitteet
2. Sisällöntuotannon perusratkaisu
3. **Sisällön syntaktinen yhteensopivuus**
4. Sisällön semanttinen yhteensopivuus
5. Annotointiprosessin kulku
6. Yhteenveto: käytäntö ja ongelmat

Syntaktinen yhteensopivuus

- Mukaan valittiin sellaisia tietoja, jotka
  - Liittyvät olennaisesti esineisiin
  - Tuovat esille eri tietoja esineistä
  - Todennäköisesti kiinnostavat museovierasta
  - Ovat todennäköisesti yhteenlinkittyneitä eri esineiden välillä
- Valitut tiedot:

MS XML kortti: ominaisuudet:

Kohdetiedot
- kohde (ArtifactType, ATW)
- materiaali (ArtifactMaterial, AMW)
- kuva
- asiasanat (ArtifactDescription, ADM)
- mitat
- kuvailu

Tekotiedot
- tekija (CreationActor, CAW)
- valmistupaikka (CreationLocation, CLW)
- valmistusaika (CreationTime, CTW)

Käyttötiedot
- kayttaja (UsageActor, UAW)
- kayttopaikka (UsageLocation , ULW)

Kokoelmatiedot
- museo
- kokoelma (MuseumCollection, MCW)

Syntaktinen yhteensopivuus (2)

- XML-muoto: Esim.

```
<esinekortti created="2004-2-17 12:32:30">
     <numero><arvo>NBA:SU5098:5</arvo>
    </numero>
     <kohde>
          <arvo>jalkineet</arvo>
          <arvo>kallokkaat</arvo>
     </kohde>
    <materiaali><arvo>turkis; poro</arvo>
     </materiaali>
    <mitat>
          <arvo>korkeus, suurin:19cm</arvo>
          <arvo>leveys, suurin:7cm</arvo>
          <arvo>pituus, suurin:26cm</arvo>
     </mitat>
     <kuvailu>
          <arvo>Tavallisen malliset kallokkaat. Kappaleet ommeltu
               useammista, epäsäännöllisistä kappaleista. Jalkineen suulla
               epäsäännöllisen levyinen, lyhytkarvaiseksi leikattu
               vyöhyke.</arvo>
    </kuvailu>
</esinekortti>
```

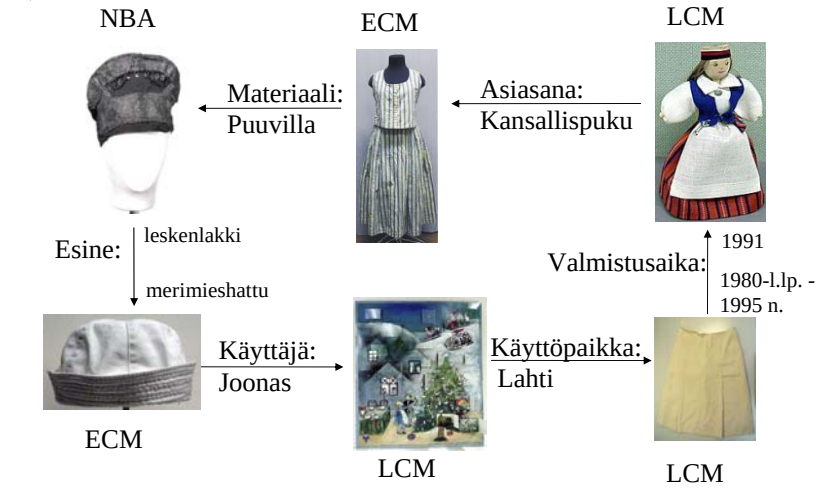
Esityksen sisältö

1. Sisällöntuotannon ongelma: lähtökohdat ja tavoitteet
2. Sisällöntuotannon perusratkaisu
3. Sisällön syntaktinen yhteensopivuus
4. **Sisällön semanttinen yhteensopivuus**
5. Annotointiprosessin kulku
6. Yhteenveto: käytäntö ja ongelmat

Semanttinen yhteensopivuus

- Semanttinen yhteensopivuus merkitsee, että:
  - Eri termistöt ovat yhteismitallisia (Esim. pikkupöydät tunnistetaan pöydiksi)
  - Termien semanttinen epävarmuus ratkaistaan (Esim. homonymia - silkki)
  - Yhteisiin resursseihin voidaan tunnistaa yksikäsitteisesti (Esim. paikat, toimijat)
- Tarkoituksena saada kokoelmien kuvailut yhteismitallisiksi sillä tavalla, että konekin ymmärtää kokoelmien väliset yhteydet ja kuvailujen sisällöt
- Semanttisen yhteensopivuuden kautta voidaan löytää ei vain sanojen, vaan myös käsitteiden ja esineiden väliset erilaiset yhteydet

Semanttinen yhteensopivuus (2)

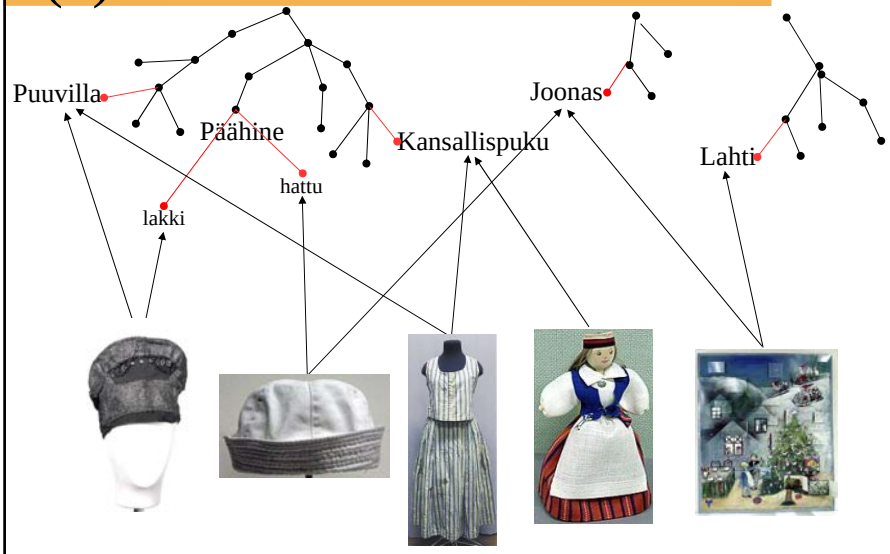


Kuvien lähteet: Espoon kaupunginmuseo (ECM), Kansallismuseo (NBA) ja Lahden kaupunginmuseo (LCM)

Rikastuttaminen ontologioilla

- Ratkaisuna on kokoelmatietojen linkittäminen ontologioihin
- Toisin sanoen sisältöä rikastutetaan ontologisilla metatiedoilla
- Yhdistäminen ontologioihin tuo:
  - Yhteismitallisuutta nimityksiin
  - Yhteisten resurssien yksikäsitteisen tunnistamisen
  - Liitoksen käsiteverkkoon. Samalla kun esine liitetään yhteen käsitteeseen, se liittyy koko käsiteverkkoon ja saa samat yhteydet muihin käsitteisiin, kuin liitoskäsitteellä jo on.
  - Yhteyden toisiin esineisiin. Esine liittyy käsitteiden kautta muihin esineisiin, jotka on jo liitetty käsiteverkkoon
  - Merkitysten ja tarkoitusperien ymmärtämisen. Ontologisilla yhteyksillä voidaan kuvata käsitteiden lisäksi merkityksiä ja tarkoituksia, kuten liittää esineet tiettyihin käyttötilanteisiin tai tarkoituksiin

Rikastuttaminen ontologioilla (2)



MuseoSuomen ontologiat

Ontologia	Sisältö	Koko
Museoalan ontologia MAO	Museoalan käsitteistö. Kehitty MASA-tesauruksesta.	n. 6900 yleiskäsitettä taksonomisesti luokiteltuna.
Esineet	Konkreettisten kokoelmaobjektien taksonomia.	MAO:n osa, 3227 luokkaa.
Materiaalit ja aineet	Taksonomia materiaaleista, joista konkreettiset objektit on valmistettu.	MAO:n osa, 364 luokkaa.
Tapahtumat	Yhteiskunnan tapahtumien ja prosessien taksonomia.	MAO:n osa, 992 luokkaa.
Toimijat	Yksilöiden, yritysten ym. toimijoiden yksilöiden ja käsitteiden ontologia.	26 käsitettä ja 1715 yksilöä.
Paikat	Paikkakäsitteiden (kylät, kaupungit jne.) ja näiden yksilöiden ontologia.	33 käsitettä ja 864 yksilöä (hyponymia ja meronymia).
Ajat	Aikakausien ontologia aikaintervalleina.	57 käsitettä.
Kokoelmat	Museoiden ja kokoelmien ontologia.	22 käsitettä ja 24 yksilöä.

Ontologiset ominaisuudet

MS XML kortti: ominaisuudet:

Kohdetiedot
- kohde (ArtifactType, ATW)
- materiaali (ArtifactMaterial, AMW)
- kuva
- asiasanat (ArtifactDescription, ADM)
- mitat
- kuvailu

Tekotiedot
- tekija (CreationActor, CAW)
- valmistupaikka (CreationLocation, CLW)
- valmistusaika (CreationTime, CTW)

Käyttötiedot
- kayttaja (UsageActor, UAW)
- kayttopaikka (UsageLocation , ULW)

Kokoelmatiedot
- museo
- kokoelma (MuseumCollection, MCW)

MS RDF kortti:
ontologioihin sidotut ominaisuudet:

Kohdetiedot
- kohde (ArtifactType, ATW)
- materiaali (ArtifcatMaterial, AMW)
- asiasanat (ArtifactDescription, ADM)

Tekotiedot
- tekija (CreationActor, CAW)
- valmistupaikka (CreationLocation (CLW)
- valmistusaika (CTW)

Käyttötiedot
- kayttaja (UsageActor, UAW)
- kayttopaikka (UsageLocation , ULW)
- kayttotilanne (UsageEvent, UEW)

Kokoelmatiedot
- kokoelma (MuseumCollection, MCW)

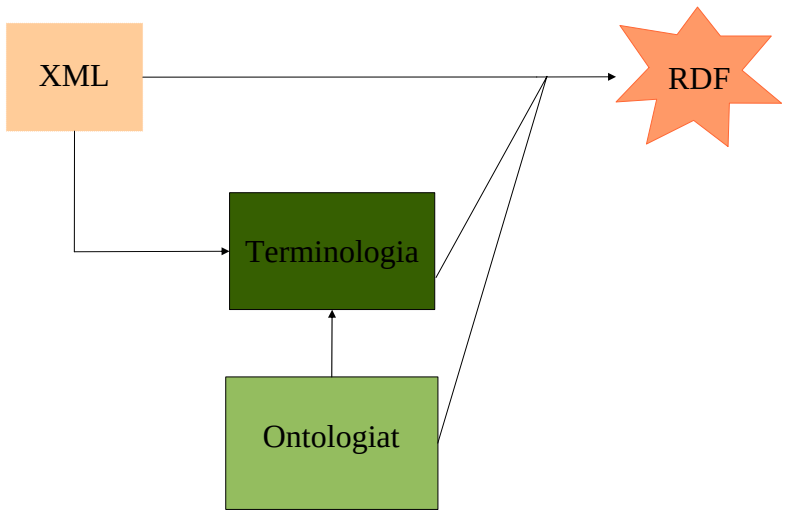
Esityksen sisältö

1. Sisällöntuotannon ongelma: lähtökohdat ja tavoitteet
2. Sisällöntuotannon perusratkaisu
3. Sisällön syntaktinen yhteensopivuus
4. Sisällön semanttinen yhteensopivuus
5. **Annotointiprosessin kulku**
6. Yhteenveto: käytäntö ja ongelmat

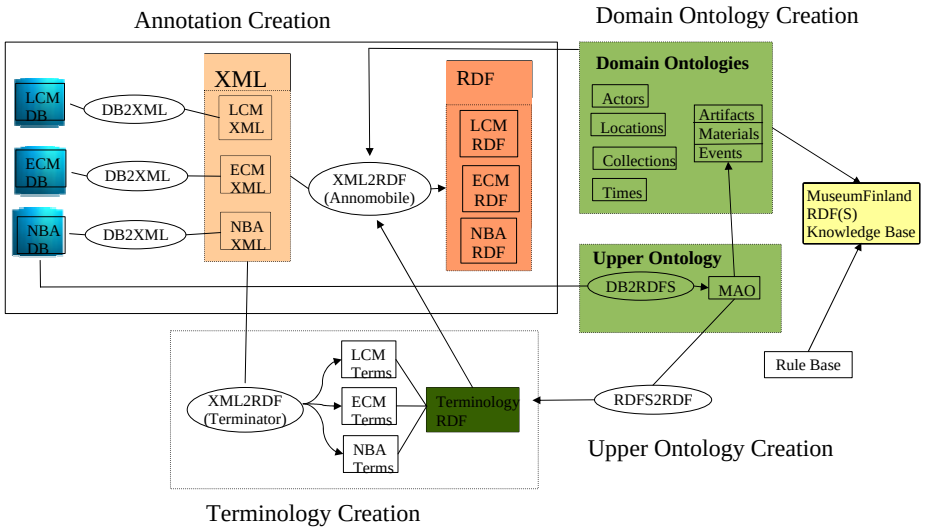
Annotointiprosessin kulku eli semanttisen metatiedon tuottaminen

- Ontologisten ripustusten luonti tehdään, kun tiedot ovat syntaktisesti yhtenäisiä
- Prosessissa tarvitaan esineiden kuvailut, ontologiat sekä ns. termiontologia
- Prosessissa kuvauskieli muuttuu monipuolisemmaksi: XML → RDF
- Prosessiin on tehty avuksi työkalut: Terminaattori ja Annomobiili

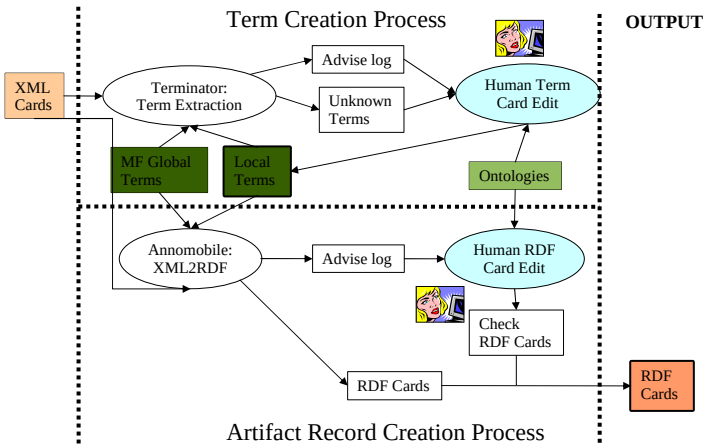
Annotointiprosessin kulku (2) - yksinkertaisimmillaan



Annotointiprosessin kulku (3)



Annotointiprosessin kulku (4) – ihmisen tehtävät osat



Esityksen sisältö

1. Sisällöntuotannon ongelma: lähtökohdat ja tavoitteet
2. Sisällöntuotannon perusratkaisu
3. Sisällön syntaktinen yhteensopivuus
4. Sisällön semanttinen yhteensopivuus
5. Annotointiprosessin kulku
6. **Yhteenveto: käytäntö ja ongelmat**

Yhteenveto: Käytännön prosessi

1. Kokoelmatietojen saaminen XML-muotoon
   - Tietojen valinta. Mitkä tiedot tietokannasta otetaan mihinkin XML-pohjan kenttään, missä järjestyksessä ja millaisessa esitysmuodossa?
   - Miten saada tiedot tietokannasta?
   - Luettelointimuodosta julkaisukelpoiseen asuun: tietojen monimuotoisuus ja korjaaminen, eri luetteloijilla erilaiset käytännöt
   - Tietojen julkaisuoikeudet
2. Termien erottaminen
   - Erottaminen hoituu automaattisesti XML-tiedoista: Terminaattori
   - Erotettujen termien annotoiminen eli ripustaminen ontologioihin
   - Mitä tehdä, jos kaikkia haluttuja käsitteitä ei löydy ontologiasta?
   - Ontologioiden täydentäminen
3. Kokoelmatietojen ripustaminen ontologioihin
   - Annotointiin on väline: Annomobiili
   - Tarkastetaan Annomobiilin tuottama tulos
   - Korjataan monimerkityksellisyydet

Päätös

- Kiitos kuuntelemisesta