# Semanttinen Web ja Web palvelut

## Sisältötuotannon ja palveluiden tulevaisuus Internetissä

Prof. Eero Hyvönen
Helsingin yliopisto ja Tietotekniikan tutkimuslaitos HIIT
Semantic Computing Research Group
http://cs.helsinki.fi/group/seco/



## Sisältö

- WWW tänään
  - Palvelut ja tiedonhaku: ongelmia
  - Tiedon esitys: merkkauskielet
  - Toiminnan esitys: ohjelmointi
- Tulevaisuuden WWW
  - Semantic Web -visio
  - Tasot ja standardit
  - Sovellusalueita
  - Demonstraatio semanttisesta portaalista



## WWW tänään



## WWW:n perusta

- URI osoitteet: resurssit
  - Sivustot, dokumentit, kuvat jne.
- HTML-kieli
  - WWW-sivujen julkaiseminen
  - Hyperlinkit
- HTTP ym. protokollat
  - Hyperteksti
  - Sähköposti
  - Keskusteluryhmät



## WWW tänään: palvelut ja tiedonhaku

- Toiminnalliset palvelut
  - Pankit, kaupat, virastot jne.
- Tiedonhaun palvelut
  - Hakukoneet (Google, AltaVista jne.)
  - Portaalit, hakemistot
    - Yahoo! jne.
    - Kallis ja hankala ylläpito
  - Arkistot, museot, kirjastot jne.
  - Tietokannat eri sovelluksissa



## Toiminnalliset palvelut: ongelmia

- Yhteisten sanastojen ja standardien puute
  - Tiedot: esim. eri yritysten tuotekatalogit
  - Prosessit: esim. ostoprosessin kuvaaminen
- Järjestelmien monimutkaisuus ja kalleus
  - Esim. EDIFACT
- Toimintojen/palveluiden monimutkaisuus
  - Esim. liikematkan osapalveluiden yhdistäminen

## Tiedonhaku: ongelmia

- Laadun mittaaminen
  - Recall: Kuinka monta % relevanteista löytyy
  - Precision: Kuinka monta % löytyneistä relevantteja
  - Relevance: Vastaako tulos haluttua
- Suomessa sanamuodot ja johdokset ongelmana



- Hakusana voi esiintyä epärelevantissa dokumentissa
  - "This page is *not* about *politics*"
- Synonyymien tunnistaminen
  - Venus =/= Aamutähti =/= Iltatähti
  - => low recall
- Homonyymien tunnistaminen
  - Nokia -> firman ja kaupungin nimi
  - => low precision



- Yleistermien käyttö vaikeaa
  - Esim: Pohjoismaat -> Suomi, Ruotsi, ...
  - Käyttäjän tunnettava eritystermit
- Fraasien käyttö
  - Esimerkiksi “Helsingin yliopisto”
- Relevanssi
  - Haun tuloksena yleensä paljon osumia
  - Dokumenttien järjestys "hyvyyden" mukaan
    - vrt. Googlen innovaatio



- Implisiittinen tieto
  - Joulu-sanalla ei välttämättä löydy pukki-sivua
- Hajautunut tieto
  - Esim. hae tutkimusryhmään kuuluvien julkaisut



## Tiedon esitys: merkkauskielet (markup languages)



## Merkkauskielten idea

- Ympäristöriippumattomia standardeja dokumenttien
  - luomiseen
  - hallitsemiseen
  - siirtämiseen
- Dokumentit tekstitiedostoja
  - Avoin yksinkertainen formaatti
  - Käytössä kaikilla HW/SW-alustoilla
  - Helppo muokata, tallentaa, lukea, siirtää
  - Käytettävissä tulevaisuudessakin

- Ideana erottaa *rakenne*, *sisältö* ja *ulkoasu*
  - Kuvataan rakenne yleisesti merkkauksilla (ohjelmoija)
    - Esim. HTML: <H1> Otsikko </H1>
  - Kuvataan sisältö (ohjelmoija)
    - Esim. XML: <OSOITE> Tietotie 3 </OSOITE>
  - Ulkoasusta päättää lukija (selain)
    - Esim. PC, kännykkä tms.



- Käytännössä työnjako menee helposti sekaisin
  - Esim. tekstin korostus tai koko on selaimen asia:
    - <EM> Korostettu teksti </EM>
      - *Loogisesti oikein; ei oteta kantaa siihen miten korostus tehdään*
    - <I> Korostettu teksti (kursiivi) </I>
      - *Loogisesti väärin, jos ajatuksena on vain korostaa tekstiä*



# HTML

- **H**yper **T**ext **M**arkup **L**anguage
- WWW-sivujen kirjoittamiskieli
- Kaikkien selaimien tukema
  - Tuettu versio kuitenkin vaihtelee!
    - Esim. HTML:n uudet ominaisuudet, Java-tuki jne.
  - Selaimet voivat näyttää sivuja hieman eri lailla
    - Esim. puutteellisten kuvausten oletusarvoiset täydentämiset, fonttivalikoimat jne.
  - Sivut on siksi aina hyvä testata eri selaimilla!



# SGML

- Standard Generalized Markup Language
- Metakieli, jonka avulla voidaan määritellä merkkauskieliä
  - Data Type Definition (DTD) -määrittelyt
  - Monimutkainen
- ISO standardi 1986
- Esimerkiksi HTML on yksi SGML:n määrittely



# XML

- E**x**tensible **M**arkup **L**anguage
- SGML:n yksinkertaisempi osajoukko (20%)
  - XML korvannut käytännössä SGML:n
- Voidaan määritellä sovelluskohtaisia markup-kieliä
  - <HENKILO>
    <NIMI> Onni Opiskelija </NIMI>
    <PUHELIN> 123456 </PUHELIN>
    </HENKILO>



- Uuden kielen syntaksin määrittely
  - DTD-kuvaus: oma määrittelykieli
  - XML Schema: määrittely XML-perustaisesti
- Ulkoasu: eXtensible Style Language (XSL)
  - XSLT: muunnokset, esim. HTML-sivuksi
  - XSL-FO: oma layout-kieli

## XML-esimerkki (DTD)

```
<?xml version="1.0"?>
<!DOCTYPE CONTACTS
[ <!ELEMENT CONTACTS   (CONTACT+)>
  <!ELEMENT CONTACT    (NAME, PHONE+, ADDRESS, IMG?)>
  <!ELEMENT NAME       (#PCDATA)>
  <!ELEMENT PHONE      (#PCDATA)>
  <!ATTLIST PHONE      type (home | work | gsm) "work">
  <!ELEMENT ADDRESS    (#PCDATA)>
  <!ELEMENT IMG        EMPTY>
  <!ATTLIST IMG        src CDATA #REQUIRED>
  <!ENTITY  Uni        "University of Helsinki"> ]>

<CONTACTS>
<CONTACT>
<NAME>Mika Klemettinen</NAME>
<PHONE type="work">191 44159</PHONE>
<PHONE type="gsm">050-54 78 595</PHONE>
<ADDRESS>&Uni;</ADDRESS>
<IMG src="mika.jpg"/>
</CONTACT>
</CONTACTS>
```



## Miksi XML?

- Samalle sisällölle eri ulkoasuja
  - Eri laitteet (PC, kännykkä, …)
  - Eri sovellukset (WWW-sivu, painettu kirja, …)
- Sisällön/rakenteen hyödyntäminen
  - Esim. parempi osumatarkkuus hakukoneissa
- Laadun kontrollointi
  - Syntaksin tarkistus mahdollista



- XML on Webin perusta jatkossa
  - Tietojen koodaus *avoimessa* muodossa
    - Runsaasti standardeja eri aloille
  - *Avoimet* rajapinnat Java ym. kieliin
    - Ohjelmallinen sivujen käsittely



## Merkkauskielten merkitys

- Muodostavat WWW:n perustan
  - Helppokäyttöisiä näyttäviä sivuja käyttäjille
  - Helppoja tehdä toteuttajan näkökulmista
  - Avoimet yhteiset standardit
- Valmistajariippumattomuus
- Stabiilisuus tiedostoformaattien muutoksia vastaan
  - Sivut ovat yksinkertaisia tekstitiedostoja
- Sovellusaluekohtaiset standardikielet



## Standardointi

- WWW-kehityksen yleiskoordinointi
  - WWW Consortium (www.w3.org)
    - Valmistajien, operaattoreiden jne. yhteistyöelin
    - Laatii WWW-suosituksia
- Sovellusaluekohtaiset organisaatiot
  - ISO: Eri alat paitsi sähkö/elektroniikka
  - IEC, CEN, UN/CEFACT, OASIS, …
  - Loputtomasti työryhmiä eri aloilla



## Toiminnan esitys: ohjelmointi

## Ohjelmointi

- WWW-selaimen sovellusohjelmointi
  - Hajautettu toiminnallisuus
- WWW-palvelimen sovellusohjelmointi
  - Keskitetty toiminnallisuus



## WWW-selaimen sovellusohjelmointi

- Java-appletit (asiakaspää)
  - Java-ohjelma siirtyy palvelimelta selaimeen
  - Ohjelma ajetaan asiakaskoneessa
- Dynamic HTML (asiakaspää)
  - ECMAScript (JavaScript, J Script)
    - HTML-koodin sekaan ajettavia ohjelmia (script)
  - Cascading Style Sheets (CSS)
    - Yleisiä tyylimäärittelyjä HTML- kielen elementeille
  - Domain Object Model (DOM)
    - Sivun oliomalli skriptejä varten



## WWW-palvelimen sovellusohjelmointi

- Server Side Includes (SSI)
  - HTML-koodilla korvattavia koodeja HTML-sivulla
    - Esim. päiväys tai muu dokumentin osa
    - Palvelin hoitaa korvaamisen ennen sivun lähettämistä



- Server Pages -systeemit (ASP, JSP)
  - HTML-sivulla myös Javaa tms. ohjelmointikieltä
  - Koodit ajetaan ja korvataan HTML-tuloksella
  - Palvelimella ohjelma luo HTML-sivut
    - Esim. tietojen haku tietokannasta
  - Tulos lähetetään selaimelle
- TAG Libraries
  - Koodit korvataan omilla merkkauksilla



- CGI-skriptit ja servletit
  - Palvelimen ohjelma
  - Saa tiedot selaimelta esim. lomakkeella
    - PUT ja GET metodit
  - Palauttaa selaimelle HTML-tuloksen



## Semantic Web & Web Services:

Visio

## Mitä hyötyä on Webistä?

- Keskeistä WWW:n tarjoamat palvelut
  - Viestintä (email, puhe, kuva, video,…)
  - Tiedonhaku (hakukoneet, portaalit,...)
  - Toimenpiteiden suorittaminen
    - Sähköinen liiketoiminta
    - Sähköinen asiointi
    - Ym.



## Kehityksen este Webissä?

- WWW-palvelu ≈ kone auttaa ihmistä
  - Edellyttää sisältöjen koneellista "ymmärtämistä"
- WWW:n sisällöt ovat ihmislukijaa varten
  - HTML, PDF, JPEG, …
- Kone ei ymmärrä WWW:n sisältöjä
  - Hakukoneet, ostoagentit, verkkomönkijät jne.
  - Periaatteessa kaikki WWW-sovellukset
- => Perustavaa laatua oleva ristiriita



## Miten Webistä tulee älykkäämpi?

- 1. Älykkäämmät sovellukset
  - Sisältö pysyy samana
  - Koneesta tehdään ihmismäisempi
- 2. Älykkäämmin esitetty sisältö
  - Sisältö helpommin ymmärrettäväksi
  - Kone pysyy tyhmempänä
- Käytännössä molempia tapoja tarvitaan
  - Yhä älykkäämmät järjestelmät käsittelevät yhä älykkäämmin esitettyjä tietoja



## Ratkaisumalli 1: Älykkäämmät sovellukset

- Kielen automaattisen tulkinnan vaikeus
  - Dokumenttien vapaamuotoisuus
  - Sisällön semantiikka
- Ei-tekstuaaliset sisällöt
  - Kuva, ääni, musiikki, video, ohjelmisto,…
  - Miten tulkita algoritmisesti?
- Tulkintaan ei riitä itse dokumentti
  - Tarvitaan konteksti, common sense
  - Tekoälyn perusvaikeuksia, ihmiselle helppoa!
- *Suuria tieteellisteknisiä haasteita*



## Ratkaisumalli 2: Älykkäämmin esitetty sisältö

- Semantic Webin lähtökohta
  - Talletetaan tieto niin, että tyhmempikin sen ymmärtää!
  - Ihminen tulee konetta vastaan
  - Kone voi auttaa ihmistä itsensä auttamisessa
- Kiihkeä kehitystyö käynnistynyt
  - W3C:n Semantic Web Activity 2001
  - W3C:n Web Services Activity 2002



## Webin sukupolvia

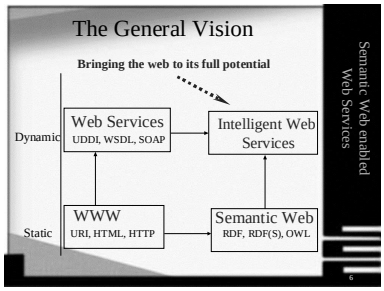
- 1G WWW:
  - WWW-sivut ihmisen tulkittavaksi
  - HTML-kieli
- 2G WWW:
  - Rakenteet ihmisen/koneen tulkittavaksi
  - XML-kieli
- 3G WWW: Semantic Web
  - Merkitykset ihmisen/koneen käytettäväksi
  - RDF(S)-kieli
- => Uusi perusta älykkäille WWW-palveluille
  - Kansainvälinen yhteistyö (W3C, ISO, FIPA, ym.…)

## WWW:n kaksi kehitysakselia

- 1. *Sisällön* rikastuminen (semantiikka)
  - Semanttinen Web
- 2. *Dynaamisuuden* lisääntyminen
  - Toiminnallisuuden lisääntyminen
    - Agent technologies, web services, grid computing
    - Kohti *aktiivisten* palveluiden verkkoa
  - Adaptiivisuuden lisääntyminen
    - Verkon rakenne ja yhteydet muuttuvat lennossa
    - Mobile systems, ambient computing



## Älykkäät verkkopalvelut




(Dieter Fensel, 2002)

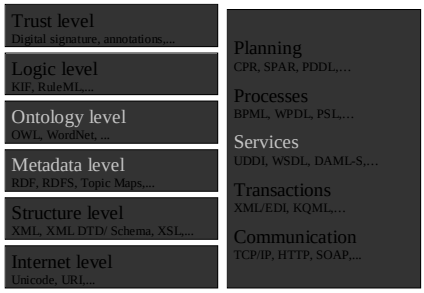


## Semantic Web & Web Services:

Teknologioita



## Semantic Web: Technology push




Laajennettu Tim Berners-Leen (W3C) ”teknologiakakku”



## Metadata level



## Miksi XML ei ole "semanttinen"?

– <OSOITE>
<NIMI>Onni Ohjelmoija</NIMI>
<PUHELIN> 123 456 </PUHELIN>
</OSOITE>

ά νΧΩΧΤωΣξ
νφΤΦΤξΧφφχ ΧϑωΩϛΥΩχωΥνΗφΤΦΤξ
νχΆϛΣυΤφξ ΘΘΙ ιΚκ νΗχΆϛΣυΤφξ
νΗΧΩΧΤωΣξ

- Semantiikka on vain nahkakansissa, ei peltikuoressa

## Semanttiset metakuvaukset

- Idea: rakenteelle on annettava merkitys (semantiikka) toisella tasolla
  - WWW-resurssien metakuvaukset
  - Käsitteiden loogiset kuvaukset
- Tärkeimmät Semantic Web -standardit
  - W3C: RDF(S)
  - ISO: Topic Map, XTM



## RDF(S)

- RDF Resource Description Framework (1999)
  - Yleinen verkkoresurssien kuvaamiskieli
  - Relaatiotietomalli, *ei* syntaksi kuten XML
- RDF Schema (2000)
  - RDF- terminologian määrittely
  - Olioajattelu WWW-kuvauksiin
    - Käsitehierarkiat, periytyminen (Class/subClass/type)



## Ontology level



## Ontologian käsite

- “Ontologia on formaali, eksplisiittinen määrittely yhteisestä käsitteistöstä” (Gruber, 1993)
  - Formaali: jämpti
  - Eksplisiittinen: konekin ymmärtää
  - Yhteinen: kommunikaatio mahdollista
- Kuvaa sovellusmaailmassa olevat käsitteet/oliot
- Ensimmäinen edellytys sille, että ihmiset ja koneet voivat ymmärtää toisiaan



## Ontologiat käytännössä

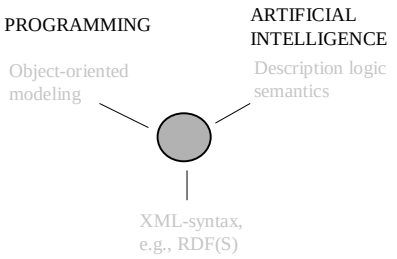
- Yleisiä hierarkkisia sanastoja
  - Esim. YSA, WordNet
- Yleisiä maailmanmalleja
  - Esim. CYC
- Spesifejä ammattitermistöjä
  - Esim. RosettaNet Dictionaries
- Luokittelujärjestelmiä
  - Esim. tuotteet/palvelut UN/SPSC
  - Esim. kirjastojen UDK
- Loogisia terminologisia malleja
  - Olioperustaisissa ohjelmistoissa



## Esimerkki: Museoalan ontologia

- Yli 6000:n käsiteen hierarkinen taksonomia
- Kehittiin HY/HIIT:n Semanttisen laskennan tutkimusryhmässä
- Sovelletaan mm. MuseoSuomi -portaalissa (Finnish Museums on the Semantic Web)

## WHAT IS NEW?






## Semantic Web & Web Services:

sovelluksia, tutkimusta



## Sovellusalueita

- Interoperability
- Informaation haku (information retrieval)
- Tietämyksen hallinta (knowledge management)
- Sähköinen liiketoiminta, Web Services
- Profilointi ja kustomointi



## Interoperability

- XML-perustaiset lukemattomat ”standardit” kommunikointikieliksi (syntaksi)
- Tiedot semanttisesti yhteismitallisiksi metakuvausten + ontologioiden avulla
- Järjestelmien yhteiskäyttö: web services
  - Avoimet WWW standardit
    - SOAP, WSDL, UDDI, WSMF, DAML-S, …
  - Legacy- ym. järjestelmien yhdistäminen web service –teknologioilla



## Informaation (täsmä)haku

- Seuraavan polven hakurobotit
  - Metatietojen hyödyntäminen
- Älykkäät hakemistot
- Matchmaker-sovellukset
- Semanttiset portaalit



## Mu land

MuseoSuomi

Demonstraatio

## What is MuseumFinland?

- A semantic portal for Finnish museums to publish their collections together on the Semantic Web
- Case study in a research project on semantic interoperability on the Web
- Duration: 3/2002-3/2004



## Research Consortium

National Board of Antiquities

Espoo City Museum

Lahti City Museum

Helsinki University Museum

ANTIKVARIA GROUP

NOKIA CONNECTING PEOPLE

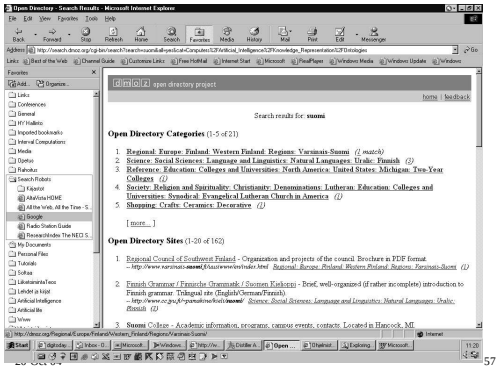
TietoEnator Building the Information Society

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

UNIVERSITY OF HELSINKI

Co-operation with: Finnish National Gallery

Open Directory Project: Koko WWW RDF(S) ontologiana!






## Case: Promoottori

- Photo repository of the Promotion Ceremonies at the University of Helsinki
- Exhibition to be opened at the Senate Square at Helsinki city center 10/2003
  - The Helsinki University Museum
- An interesting view into the life of the university
- Complicated semantics of obscure events
- Novice users who do not know the contents before
- Team: Eero Hyvönen, Samppa Saarela, Kim Viljanen, Arvil Styrman, Jaana Tegelberg, …



## Tietämyksen hallinta (knowledge management)

- Ongelmia
  - Dokumenttien monimuotoisuus
  - Maapalloistuminen -> sisältöjen hajautus
  - Tietämyskannan komplisoituminen
- SemWeb-tekniikat antavat uusia työkaluja
  - Liima heterogeenisten hajautettujen dokumenttien hallintaan
- Adoben XMP
  - Kaikki WWW-julkaisut tukevat RDF-metadataa



## Sähköinen kaupankäynti: Web Services

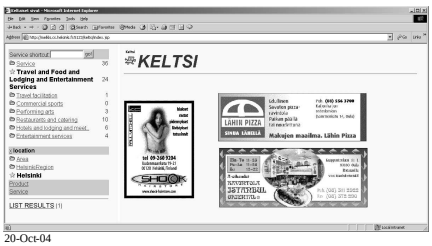
- Miten tarjoan oman tuotteen/palvelun kansainvälisille markkinoille?
  - Visio: globaalit rekisterit ja sanastot käytettävissä (UDDI-rekisterit, RosettaNet,...)
- Miten hoidan transaktiot ja prosessit kumppaneiden kanssa?
  - Esim. tarjouspyyntöön vastaaminen
  - EDI-XML, SOAP, WSDL
- Toimintakehykset: ebXML, …

- Semantic Webin mahdollisuuksia
  - Sisältöjen ja tietojärjestelmien yhteentoimivuus
  - Eri standardien yhdistäminen
    - Esim. toimialaportaalit
  - Sisältöjen rikastaminen
  - Liiketoimintaprosessien automatisointi
    - Semanttset kuvaukset



## Case: älykkäät Keltaiset sivut

- Ontologiat haun perustana
- Toisiaan tukevien palveluiden suositus






## Adaptiivisuus: profilointi ja kustomointi

- Ihmiset ja palvelut (P3P)
  - Omat preferenssit, tietosuoja,...
- Laitteet (CC/PP, FIPA Device Ontology)
  - Esim. MV-kännykälle ei värikuvia
- Dokumentit (transcoding)
  - Sisältöjen kustomointi eri laitteille ja tarpeisiin



## Semantic Web & Web Services

- Juna lähti jo
  - XML-standardointihankkeet, 90-luvun loppu
  - W3C Semantic Web Activity , 2001/kevät
  - EU:n OntoWeb 2001/kesä
  - W3C Web Services Activity, 2002/kevät
- Suomessa tapahtunutta
  - “Semantic Web Kick-Off in Finland”, 2001/syksy
  - “Towards the Semantic Web and Web Services”, 2002/syksy
  - Tim Berners-Lee saa Millenium-palkinnon 2004/kesä
  - “Web Intelligence – Älyä verkossa”, 2004/syksy
- Tärkeä ala jatkossa monessa mielessä
  - Teollinen intressi
  - Tekninen mahdollisuus
  - Tieteellinen haaste
  - Kansallinen intressi