8. Digitaaliset kirjastot

· keskeinen tiedonhaun ja tiedon jakelun kehityskohde:

- konferensseja 1994 ..
- teemanumeroita:
  Comm. ACM 38,4 (1995); 41,4 (1998)
  IEEE Computer 29,5 (1996); 32,2 (1999)
  SIGLINK Newsletter 4,2 (Sept. 1995)
  SIGOIS Bulletin 16,2 (Dec. 1995)
- D-Lib Magazine

· mallina periaatteessa perinteinen kirjasto
  erilaisia painotuksia
  skaalaeroja

"digitaalisessa muodossa olevien tietojen kokoelma"

- hajautettu, usein laaja
- sovellusalakohtainen, organisaatiokohtainen
- maailmanlaajuinen
- ainakin perinteisen kirjaston toiminnot
  "yksi monimutkaisimpia ja edistyneimpiä informaatiojärjestelmiä"

· ominaisuuksia
  materiaali enimmäkseen digitaalista
  haku ja käyttö tietokoneavusteista
  ylläpito, luokittelu pitkälle automatisoitu
  jokin hallinnollinen tai sisällöllinen rajaus

· elementtejä, tietämystä eri aloilta:
  - tiedonhaku, suodatus
  - dokumenttien tallennus, arkistointi
  - informaation selektiivinen jakelu
  - hajautetut tietokannat
  - yhteistyön tuki
  - erilaisia esitysmuotoja, multimedia
  - osien yhdistäminen, heterogeenisuuden hallinta
  - resurssien hallinta
  - hyperteksti
  - tekijänoikeudet

· suhde WWW-materiaaliin?
  WWW luonnollinen toteutusalusta
  WWW:ssä myös vapaamuotoista,
  tilapäistä materiaalia
  dokumenteissa usein viittauksia
  toisiin dokumentteihin
  → monimutkaistuu!

Digitaalinen objekti: yleisnimi digitaalisen kirjaston 'perusalkiolle'
- tallenne, raportti, dokumentti, äänite
  myös interaktiivinen, ohjelma

- objektille tarvitaan yksikäsitteinen tunniste
  vrt. ISBN

- Internet: URL → URN, URI
  (ei paikkasidonnainen)

- periaatteessa kahva (handle) ja

  n / t, missä

n viittaa kahvapalvelimeen
t on yksikäsitteinen (palvelinkohtainen)

- kahvapalvelin tuntee t:tä vastaavan
  todellisen osoitteen

Esim. handle.resolver.org/minun.sivu →
http://www.cs.helsinki.fi/u/erkio

8.1 Metadata ja sen käyttö

metadata = digitaalisen objektin ominaisuuksia
kuvaileva tieto ("tietoa tiedosta")

- standardeja kehitetty
  · perinteinen MARC-formaatti
  · Dublin Core-määrittely
  · RDF (Resource Description Framework)
  XML-pohjainen esitystapa

Dublin Core: semanttinen kuvaus

- WWW-resurssien määrittelyyn
- myös kirjastot, museot jne
- yhteys viralliseen
  WWW-metatietoarkkitehtuuriin (W3C)

- tavoitteita:
  yksinkertaisuus (maallikoillekin)
  selkeä ja yhteisesti sovittu semantiikka
  kansainvälinen
  laajennettavissa
  (tarkenteet, omat elementit)
  eri hakujärjestelmien yhdistämisen
  edistäminen
- syntaksi HTML:n META-kenttään sijoittamista
  varten, myöhemmin RDF-esitys

Dublin Core - elementit:

1. Title objektin nimi, otsikko
   Title = The Warwick Metadata Workshop

2. Subject objektin käsittelemä aihe
   Subject = Description of electronic resources

3. Creator objektin tekijä (kirjoittaja, kuvaaja, ...)
   Creator = (scheme=Personal)
   Dempsey, L. & Weibel, S.L.

4. Description kuvaus tekstinä
   esim. tekstidokumentin abstrakti

5. Publisher julkaisija
   Publisher = Addison-Wesley (HY/TKTL)

6. Contributors muut objektin valmistamiseen osallistuneet
   Contributors = Kalle Kuvaaja, Maija Kääntäjä

7. Date (julkaisu)päivä
   - tarkenteita:
     Created = (alkuperäinen) luontiaika
     Issued = julkaisuaika (esim. versio CD)
     Accepted (kun tarvitaan)
     Available käytettävissä
     Date (Scheme = Available)
     1999-01-16..1999-05-31
     Acquired hankintapäivä
     Valid voimassa
     esim. aikataulu tms. faktajulkaisu

8. Type tyyppi, laji
   Esim. kotisivu, tekninen raportti, sanakirja

9. Format muoto
   teknisen hyödyntämisen kannalta
   Format = text/html

10. Identifier tunniste
    Esim. *Identifier = (Scheme = ISSN) 1082-9873 =*
    (Scheme = URL) http://www.dlib.org/dlib/a1.htm

11. Source lähde
    - tavallaan toisen objektin tietoa !
    Source (Scheme = ISBN) = 0-210-54321-X

12. Language kieli
    Language = English

13. Relation suhde toiseen dokumenttiin
    (ei vielä lopullisesti hyväksytty?)

    - IsPartOf & HasPart kokonaisuus, osa
    - IsVersionOf & HasVersion
    - IsFormatOf & HasFormat
      sama sisältö, eri muoto
    - References & IsReferencedBy
      viittaus, lainaus
    - IsBasedOn & IsBasisFor
      esim. käännös, sovitus
    - Requires & IsRequiredBy
      käyttö vaatii toista objektia

14. Coverage kate
    - ajallinen t. maantieteellinen kate
      Coverage (type = location) = Europe
      Coverage (type = temporal, scheme = free)
      = 20th century
      Coverage (Scheme = temporal)
      = (150796..311296)

15. Rights tekijänoikeus
    Right = "public domain"
    Right (Scheme = URL) = .../copyright.html

8.2 Rakenteeseen perustuvasta hausta

Esim. XML-dokumentti:
dokumentin teksti jakaantuu rakenneosiin

- haku voidaan kohdistaa osiin:
  'semistructured' in title

- termien lähekkäisyys häviää
- dokumentin rakenteellisuus / tekstin eteneminen

```
Esim.
  <Publication URL ="ftp://db.stanford.edu/xml.ps"
        Authors="RG JM JW">
    <Title>From semistructured Data to XML: ...</Title>
    <Published>Proceedings of WebDB'99</Published>
    <Pages>25-30</Pages>
    <Location>
        <City>Philadelphia</City>
        <State>Pennsylvania</State></Location>
    <Date> ....</Date>
  </Publication>
  <Publication URL="....." Authors="TL SA JW">
    <Title>Integrating structured and semistructured data</Title>
    <Institution>Stanford Db Group</Institution>
    . . .
  </Publication>
  <Author ID="SA">S. Abiteboul</Author>
  <Author ID="RG">R. Goldman</Author>
    . . .
```

Rakenteisten dokumenttien käsittely kehittyy
- kyselykieliä on esitetty
- yhteys tietokantaan (SQL ...), tiedonhakuun

Esimerkkejä operaatioista
(Baeza-Yates & Navarro, 2000):

- dokumentin osien tunnistus
chapter kaikki luvut

- rakenteen mukaiset määreet
citation in table ei kaikki sitaatit
title child chapter vain lukujen otsikot
<chapter> . . . .
<title> . . . </title> . . .</chapter>
chapter parent(3) section

- sanojen esiintymät rakenteissa
section with(5) “computer” aliluvuissa
vähintään 5 esiintymää

- järjestyssuhteet, etäisyys

table after figure (in chapter)
samassa luvussa
“computer” before(10) “architecture” (paragraph)
enintään 10 symbolin etäisyydellä,
samassa kappaleessa

8.3 Dokumenttikopioiden havaitseminen

- digitaalisen kirjaston heterogeeninen luonne → kopioita tai osittaisia kopioita on olemassa

- monia tarpeita:

  - tekijänoikeuksien valvonta, maksun periminen
  - haun tuloksen “minimointi”: duplikaattien poisto
  - turhan käsittelyn (lukemisen) välttäminen

- eri asteita:
  täydellinen kopio
  osittainen kopio (myös IsPartOf jne)
  sama sisältö, eri esitysmuodot

- hallinnollinen ongelma
  objektien rekisteröinti

- tekninen ongelma
  kopioiden tunnistaminen,
  samanlaisuuden määrän havaitseminen

- periaatteellinen ongelma
  halukkuus antaa materiaalia kirjastoon

- taloudellinen ongelma
  korvaus hyväksikäytöstä

Stanfordin COPS-projekti:

- jaetaan dokumentti sopiviin yksiköihin, esim. lauseisiin
- lasketaan dokumentille hajautukseen perustuva tunniste, esim. yksiköittäin
- lasketaan ‘uudelle’ dokumentille tunniste samalla periaatteella ja verrataan tunnistetta vanhoihin
- kynnysarvon ylittävä samankaltaisuus voi merkitä kopiointia

- useita suunnitteluparametreja
  yksikön koko: lause, kappale, ...?
  lasketaanko koko dokumentista vai vain osasta?
  otetaanko yksikköjen järjestys huomioon?
  yksikköjen erottaminen eri tiedostomuodoista
  Word, LaTex, PostScript, ...

  yksikköjen kokoaminen tarkasteluun eri tavoin:
  jokaiselle yksikölle oma hajautus
  A, B, C, D, E, F
  hajautusjakso = k erillistä yksikköä
  ABC, DEF
  hajautusjakso (k) päällekkäinen (esim. k-1 yksikön osalta): ABC, BCD, CDE, DEF
  hajautusjakson pituus määrätään hajauttimella (vaihtelee) esim. AB, CDEF

SCAM-menetelmä (Shivakumar, 1999)

- pyritään huomaamaan dokumentin sisältyvyys toiseen → epäsymmetrinen mitta

- kosinimitan tulokset huonoja, kun termien frekvenssit kovin erilaisia

- määritellään termien läheisyysjoukko c(d1,d2) = ne termit, joiden frekvenssien ero on kohtuullinen, esim. enintään kaksinkertainen

- lasketaan vain lähijoukkoon kuuluvien termien pistetulo, normeeraus toisen dokumentin kaikilla termeillä

- termit voitaisiin myös painottaa

- kokeissa saatu parempia tuloksia kuin kosinimitalla (vastaavuus intuitiiviseen samanlaisuuteen)

Muita menetelmiä:

- merkkijonopohjaiset

  hajautus kaikille w:n peräkkäisen sanan jonoille,
  saadaan dokumentin nimikirjoitus;
  nimikirjoitusten vertailu

  (lauseita ei tarvitse tunnistaa)

  w:n valinta?
  - liian pieni: yksittäiset yhteiset sanat dominoivat
  - liian suuri: erilaisuus ei näy

- suffiksipuihin perustuva matchDetectReveal (harj.)

Hypertekstimuotoisten dokumenttien samanlaisuus
(kopioiden mielessä)

- tekstiin perustuva
- rakenteeseen (topologiaan) perustuva

- teksti voidaan linearisoida, jos solmuilla on
  yksikäsitteinen järjestys
  - rakennelinkit / muut
  - puurakenne, jono; verkko

- osittainen linearisointi (vastaa tavallaan sanajonoja):

- muodostetaan linkkirakenteen mukaisesti rypäitä,
  esim. “yhden linkin yhdistämät” (solmuparit)
  “enintään kahden linkin yhdistämät solmujonot”
  “isäsolmu ja kaikki lapset”

  vrt. hierarkkinen dokumentti:
  luvut peräkkäisinä lapsisolmuina

- voidaan laskea
  - vastinrypäiden suhteellinen osuus
  (tietyn samanlaisuustason ylittävät)
  - jokin koostesuure (keskiarvo) kaikista tai
  jollain tavalla valituista (ryväs/ryväs)-
  samanlaisuusarvoista

Samanlaisuusmittojen käyttö:

- dokumenttien kopiot (plagiaatit)
- dokumenttien versiot
- mirror-palvelimet (sisältöjen vertailu)