Luetteloinnin ja tiedonhaun tulevaisuuden näkymiä

Juha Hakala
Helsingin yliopiston kirjasto
2006-03-23

Esityksen rakenne

- Johdanto
- Portaalien metadata
  - Kokoelmien kuvailu (kokoelmakartta)
  - Tiedonhakupalvelujen kuvailu
- Digitaalisen aineiston erikoispiirteitä
  - Pitkäaikaissäilytyksen metatiedot, automaattinen indeksointi
- Perinteisen luetteloinnin muutokset
  - Formaatit; kopioluettelointi, teosten kuvailu

Johdanto

- Luetteloinnin tavoitteet ja toteutustapa riippuvat muun muassa:
  - Kuvailun kohteista
  - Käytössä olevista järjestelmistä
  - Henkilö- ja muista resursseista
- ”Triangeli” lisää sekä sovellusten määrää että kuvailun kohteita, mutta prosessi on ollut osin hallitsematon: ensin tulevat välineet (ohjelmisto), sitten aletaan miettiä kuvailusääntöjä
- Ammattitaitoisia luetteloijia tarvitaan edelleen
  - Korkeatasoiset luettelot ovat keskeinen kilpailutekijä

Portaalien metadata

- Perinteisen kirjastojärjestelmän tavoin portaali tarvitsee korkealaatuisia kuvailutietoja toimiakseen kunnolla
- Kuvailun kohteet vaihtuvat: niitä ovat kokoelmat ja Internet-tietokannat, eivät enää julkaisut
- Samalla katoavat luettelointisääntöjen ja MARC-formaatin tarjoama vakaa perusta
- Portaalien metadata on ollut laadultaan huonoa ja heikosti vaihdettavissa järjestelmien kesken

Kokoelmien kuvailu: yleistä

- Tavoitteena auttaa tiedonhakijaa paikallistamaan relevantteja kokoelmia kohdetietokannoista
- Välttämätön välivaihe tiedonhaussa viimeistään silloin kun portaalissa on tuhansia tietokantoja
  - Nelli: nyt 500; määrä kasvaa jatkuvasti
- Kaksiportaisen tiedonhaun toteuttaminen (kokoelmat – julkaisut) on suurin haaste
- Perinteinen kirjastojärjestelmä & MARC eivät sovellu kokoelmien kuvailuun (toistaiseksi)

Kokoelmien kuvailun historiaa

- Alkujuuret USA:ssa: Conspectuksen käytöllä pyrittiin kokoelmien aihealueiden ja vahvuuksien määrittelyyn
- Ristiriitaisia tuloksia kuvailun objektiivisuudesta – eri kirjastojen yhteismitallisuus riittämätön
- Integrointi perinteiseen kirjastojärjestelmään heikkoa

Kokoelmien kuvailun historiaa (2)

- Kokoelmien kuvailun renessanssi käynnistyi Englannista 90-luvun lopulla
- RSLP:n kokoelmien kuvailuhanke
  - http://www.ukoln.ac.uk/metadata/rslp/
- Tavoitteena kokoelmien kuvailu johdonmukaisesti ja konelukuisella tavalla
- Laadittiin analyyttinen malli kokoelmista ja niiden luetteloista
  - http://www.ukoln.ac.uk/metadata/rslp/model/

Kokoelmien metadataformaatit

- RSLP kehitti kokoelmien kuvailuformaatin, josta tuli kaikkien myöhempien määritysten kantaäiti
- RSLP:n perillisiä ovat sekä Dublin Coren Collection Description –profiili sekä NISO:n Metasearch Initiativen Collection Description Specification (Z39.91 -standardi)
  - Useimmat hankkeet soveltavat jotakin näistä formaateista; jatkossa Z39.91 valtaa alaa
- UKOLNilla keskeinen rooli kehitystyössä (Andy Powell & Pete Johnston päävastuussa)

Z39.91 Collection description specification

- DC Collection Description Application Profilen laajennus; lisätty esimerkiksi mahdollisuus kuvata kokoelman kattavuus
- Sisältää DC:n 15 peruskenttää sekä 13 kokoelmien kuvailun erikoiskenttää, toisin sanoen vain oleellisen (RSLP laajempi)
  - Odotettavissa on että formaatti kasvaa käytön myötä MARCin tavoin (20 -> 200 kenttää)
- Määrittelee XML-pohjaisen vaihtoformaatin
- Saatavissa osoitteesta: http://www.niso.org/standards/resources/Z39-91-DSFTU.pdf

Z39.91 (2)

- Formaatin luonnos julkistettiin marraskuussa 2005, lopullinen versio tulossa loka-marraskuussa 2006 pienin muutoksin ellei yllätyksiä satu
- Käytännön sovelluksia toistaiseksi vähän; niistä yksi on oma Kokoelmakartta-hankkeemme
- Täyttä yksimielisyyttä kokoelmien kuvailun merkityksestä ei vielä ole, vaikka Englannissa se on jo ”normaalia” toimintaa

Kokoelmien kuvailun haasteita

- Luettelointisääntöjä ei ole, ja erimielisyys vallitsee siitä, ovatko ne edes tarpeen
  - Toistaiseksi selvittävä projektien omilla oppailla
- Joidenkin tietojen järkevä tallennus vaikeaa
  - Kokoelmien ID-tunnukselle on kenttä, mutta ei standardia. Itse keksitystä tunnuksesta voi olla haittaa tietoja vaihdettaessa
- Tietojen vaihdon periaatteet
  - Sovelletaanko OAI-PMH:ta, ja jos niin miten?
- Mistä löydämme tallentajaresurssit?
  - Kokoelmien kuvailua ei voi automatisoida

Kokoelmatietojen keruun haasteita

- Miten laajalti kokoelmien kuvailuja pitäisi haravoida?
  - Vain parasta / suuri on kaunista / omalta erikoisalalta kaikki mahdollinen
- Keruuprotokolla (OAI-PMH) asettaa tietyt rajat sille, miten valikoivia voimme olla
- Miten saada tallennus kansainväliseksi (monikieliset kuvailut) alusta lähtien
  - Emme ole tottuneet siihen, että kuka tahansa voi kerätä metatietojamme milloin vain minne vain

Internet-tietokantojen (palveluiden) kuvailu

- Tavoite kuvata verkon kautta käytettävissä olevan tietokannan hakuliittymä niin, että portaali tai muu sovellus voi tehdä siitä tehokkaasti hakuja
- Kuvaus tehdään ”konetta varten” ja se voidaan rakentaa ohjelmiston avulla, jos tietokanta on haettavissa jonkin standardin nojalla
- Ihmisvoimin kuvailun teko voi viedä hyvin paljon aikaa, ja tulos voi olla silti puutteellinen (ja vanhentua nopeasti)
  - Epästandardit tietokannat päänsärky portaalien ylläpitäjille

Palvelujen kuvailu: Z39.92

- Alussa oli Z39.50 Explain-palvelu, 90-luvun puolivälissä
  - Teknisesti hyvin monimutkainen, ei toteutuksia
- Explain Lite kehitettiin 90-luvun lopulla ONE-2 –projektissa; erona edeltäjään oleellisesti helpompi (vaan ei tarpeeksi helppo) toteutus
- ZeeRex-määritys osa Z39.50:n modernisointia SRU/SRW-standardeiksi; pohjana Explain Lite
- ZeeRex päivitetty NISO Z39.92 –standardiksi (Information Retrieval Service Specification)
  - Sisältää tiedonhakupalvelujen kuvaamiseen tarvittavat tietoelementit sekä datan vaihtomuodon

Palvelujen kuvailu: sovellukset

- Index Data ja HYK päässeet maaliskuussa 2006 yksimielisyyteen hankkeesta, jossa kehitetään open source –sovellus palvelukuvausten automaattista generointia ja OAI-PMH -vaihtoa varten
- Tavoitteena on, että Ex Libris, Endeavor ja muut ohjelmistotoimittajat integroivat sovelluksen omiin tuotteisiinsa
  - Portaalien metatietojen määrän kasvattaminen sekä tietokantojen tietojen vanhenemisen välttäminen

Palvelujen tietojen keruun haasteita

- Miten suuriksi haluamme portaalimme kasvattaa
  - BookWheressä noin 2000 kohdetietokantaa; kopioluetteloinnissa sekään ei ole tarpeeksi, mutta miten paljon tiedonhakijat tarvitsevat
- Jos kaikki keräävät tietokantojen tietoja, miten paljon lisäkuormaa siitä tulee?
- Mitä teemme epästandardeille järjestelmille?
- Miten saamme tietoa tietokantojen sisällöistä?
  - Palvelukuvausten koplaaminen kokoelmien tietoihin
- Semanttinen yhteismitallisuus: jos sitä ei ole, ei yhteishakukaan onnistu

Digitaalisen aineiston erityispiirteitä

- Periaatteessa tiedämme miten digitaalista aineistoa luetteloidaan; ISBD(ER) määrittelee miten homma hoituu, ja formaatista löytyvät tarvittavat kentät
- On kuitenkin perusteltuja syitä epäillä, ettemme tiedä asiasta vielä tarpeeksi
- Esimerkki: pitkäaikaissäilytyksen metadata

Pitkäaikaissäilytyksen metadata

- Kuvailutiedot, joiden avulla tallenne voidaan säilyttää käyttökelpoisena
- Tietojen sisältö riippuu valitusta säilytysstrategiasta
  - Migraatio (dokumentin konvertointi) edellyttää eri tiedot kuin emulointi (käyttöympäristön jäljittely)
- Metatieto standardi puuttuu, lupaavin ehdokas Premis-työryhmön ehdotus
  - http://www.oclc.org/research/projects/pmwg
- Varsinaiseen standardiin vielä paljon matkaa

Automaattinen indeksointi

- ”Graalin malja”, tekeillä jo 70/80-luvulta
- Joissakin kirjastoissa (British Library) tavoitteena perinteisen luetteloinnin osittainen korvaaminen
- HYK:ssa pyritään saamaan haettavaksi aineisto jota ei voida luetteloida (verkkojulkaisut, digitoidut sanomalehtiartikkelit)
- Laadun avain lähtötekstin rakenteisuus /rakenteistaminen
  - Mikkelin tuotantoprosessi vs. Arto: 10-1 nopeudessa, laadusta meillä ei vielä ole kokemusta

Kopioluettelointi

- Edellyttää uuden organisaation (hankinta ja luettelointi yhteen) ja työmenetelmät ja taidot hyvien välineiden (Z39.50-sovellukset ja formaattikonvertterit) lisäksi
- Tehokas soveltaminen toistaiseksi epätavallista, kansainvälisestikin
  - Poimitaan harvoista kohdetietokannoista, yhdestä kannasta kerrallaan, editoidaan tietueita käsin, tehdään tarpeettomia muutoksia (luettelointikieli)

Suomen malli

- Tehokas Z39.50-asiakasohjelma, joka sisältää paljon kohdetietokantojen tietoja
- Usemarcon-formaattikonvertteri, jolla on rakennettu kopioijille sopivia konversioita, viimeksi kyrilliikan luettelointiin
- Sopivien kohdetietokantojen lisensointia sekä luetteloijien välistä tiedonvaihtoa kokemuksista
- Joustoa luetteloinnin periaatteisiin

Kopioluetteloinnin vaikutuksista

- Ammattitaitovaatimukset eivät laske, mutta ne muuttuvat erilaisiksi
  - Tiedettävä mistä tietuetta kannattaa hakea
  - Kyettävä näkemään, mitä pitää muuttaa
  - Osattava vaatia parannuksia välineisiin; esimerkiksi konversiotauluihin
- Oleellisin vaikutus tietueiden laadun paraneminen; työskentelyn nopeutuminen vasta toisella sijalla (ja se toteutuu vain jos välineitä käytetään taitavasti ja organisaatio on sopeutettu)

Kopioinnin keskitetty tuki

- Lisää maksuttomia kohdetietokantoja
  - Oravannahkakauppaa, myyntiartikkelina Fennica, Viola, Linda etc.
- Lisää tietoa siitä, mistä mitäkin saa
  - Kokemuksen siirto luetteloijien välillä
- Paremmat välineet
  - Lisää ja parempia Usemarcon-konversioita
  - Lisää kohdetietokantoja BookWhereen
- Järjestelmätoimittajien painostaminen
- Standardien kehittäminen

Formaatit: MARC

- MARC21:n asema vahvistumassa kansainvälisesti; Suomi osa tätä trendiä
  - 2008 MARC21 käyttöön Voyager-kirjastoissa; 2006 päätetään otetaanko MARC21 kansalliseksi formaatiksi
- MARC21 ei ole teknisesti paras formaatti, mutta tarpeeksi hyvä, ja sen käyttäminen tekee tietojen ja järjestelmien vaihdon helpommaksi
- Vaihtomuodoksi MarcXchange ISO2709:n rinnalle
  - MARCXML:n formaattiriippumaton laajennus, paljon pieniä muutoksia siihen verrattuna

Formaatit: Dublin Core

- 10 vuoden iässä DC on teknisesti ja organisatorisesti suhteellisen vakaa
  - Pitkällä aikavälillä ylläpito-organisaatio auki
- Perus-DC:n asemesta käytetään sovellusprofiileja (asiakirjat, kokoelmat) tai hankekohtaisia laajennuksia
- Suomessa kansalliskirjasto tukee käyttöä ja uusien sovellusprofiilien tekemistä

Formaatit: ONIX

- Kustantajat kehittivät oman formaatin, kun eivät kyenneet muuttamaan Dublin Corea mieleisekseen
- Kirja-alan ratkaisuna hiipii myös kirjastojen agendalle (uusi ISBN edellyttää ONIXia)
- Keskeinen ongelma koulutuksen ja tuen puute
  - Vaikka formaatti sinänsä olisikin OK, sillä tuotettu data on yleensä heikkotasoista
- EditEUR on toivonut, että Suomeen perustettaisiin ONIX-tukiyksikkö HYK:oon
  - Mahdollisesti yksi Kirjan Tie –hankkeelle tarjottava palvelu; perustamiskulut?

Teosten kuvailu

- Teoriatasolla eri koulukuntia (kirjastot / kustantajat) jotka hahmottavat asiaa eri näkökulmista ja käsittein
  - FRBR versus EditEUR-malli
- Sääntötasolla analyysia teos/manifestaatiotason elementeistä
- Järjestelmissä ”fuskataan” tai ei tehdä mitään
  - FRBR-toteutus käyttöliittymässä; käyttäjä näkee teostiedot vaikka niitä ei ole oikeasti tietokannassa
  - Virtua, uusi WorldCAT, library.dk
  - Tämäkin edellyttää uutta järjestelmäarkkitehtuuria

Teosten kuvailu ja Voyager

- Endeavorin kanssa on yritetty aloittaa keskustelua, huonolla menestyksellä
- Vaikeaa arvioida, kuinka vaikeaa FRBR:n toteuttaminen WebVoyage’ssa / tietokantatasolla olisi
  - Veikkaus: ottaen huomioon miten Voyager tallentaa bib. datan, FRBR:n implementointi on vaikeaa, kuten useimmissa muissakin kirjastojärjestelmissä
  - Virtuassa sisäinen tallennusmuoto on XML, minkä vuoksi eri tulostusmuotoja on helppo rakentaa