Protokoli za distribuirano poizvedovanje po bibliografskih zbirkah

Z39.50, OAI-PMH in SRU

Viktor Harej
viktor.harej@ff.uni-lj.si

Predstavitev podatkov v računalniku

- Podatki in ukazi so v računalnku predstavljeni **digitalno (diskretno)** - s pomočjo električne napetosti. Računalnik na strojnem nivoju razlikuje le dve napetostni stanji: 0, 1
- **Podatek, ki lahko zavzema samo dve vrednosti, od katerih sta obe enako verjetni, je bit**. (BInary digiT).

Dogovor na najnižjem nivoju

- Oblika oz. predstavitev, v kateri je informacija shranjena, je odvisna od njenega pomena:
  - Predstavitev števil
  - Predstavitev znakov

  http://sl.wikipedia.org/wiki/ASCII

Komunikacijski protokol

- je **formalni opis sporočil v elektronskem formatu in pravil za izmenjevanje teh sporočil znotraj ali med posameznimi računalniškimi in telekomunikacijskimi sistemi**.
- lahko vključuje signaliziranje, avtentikacijo, opis napak in popravljanje napak.
- določa sintakso, semantiko in sinhronizacijo komunikacije in je lahko implementiran znotraj programske ali strojne opreme (vir: Wikipedia)
- analogija z diplomatskim protokolom
- primer: protokol http

Distribuirano poizvedovanje

- Iskati in pridobiti informacije (zapise) v oddaljeni podatkovni bazi
- Distribuiran sistem je sestavlje iz več med seboj avtonomno delujočih sistemov, ki med seboj komunicirajo preko računalniškega omrežja
- Distribuirano poizvedovanje: iskanje in pridobivanje informacij po distribuiranih sistemih

Protokoli za distribuirano poizvedovanje

- Z39.50 in OAI-PMH standarda
- SRU specifikacija
- V prvi vrsti namenjeni implementatorjem – proizvajalcem programske opreme
- Glavni razlog je interoperabilnost

Z39.50

- Z39.50 je aplikacijski protokol za komunikacijo med računalniškimi sistemi, prvotno namenjen rabi v knjižnicah in ostalih institucijah, povezanimi z iskanjem informacij
- Tekom Z-zveze klient in strežnik izmenjata serijo sporočil
- Dinamični protokol

Z39.50

- Sinhronizacija
  - Definirana v standardu
- Semantika
  - Definirana v standardu
- Sintaksa:
  - Definirana v priponki z ASN.1/BER jezikom

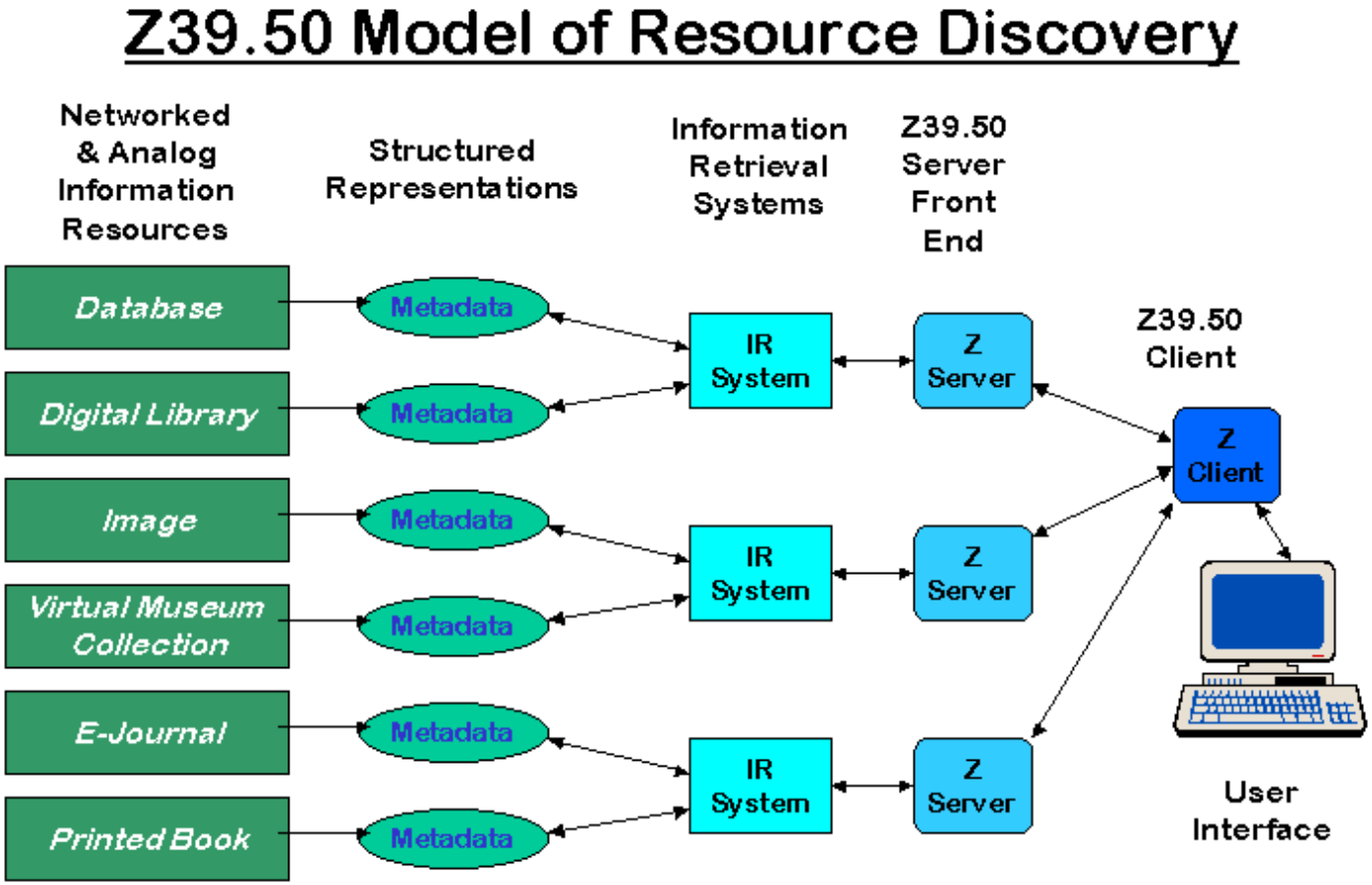
ASN.1

- Mednarodno standardiziran, platformsko neodvisen in jezikovno neodvisna notacija za specificiranje podatkovnih struktur
- http://www.loc.gov/z3950/agency/Z39-50-2003

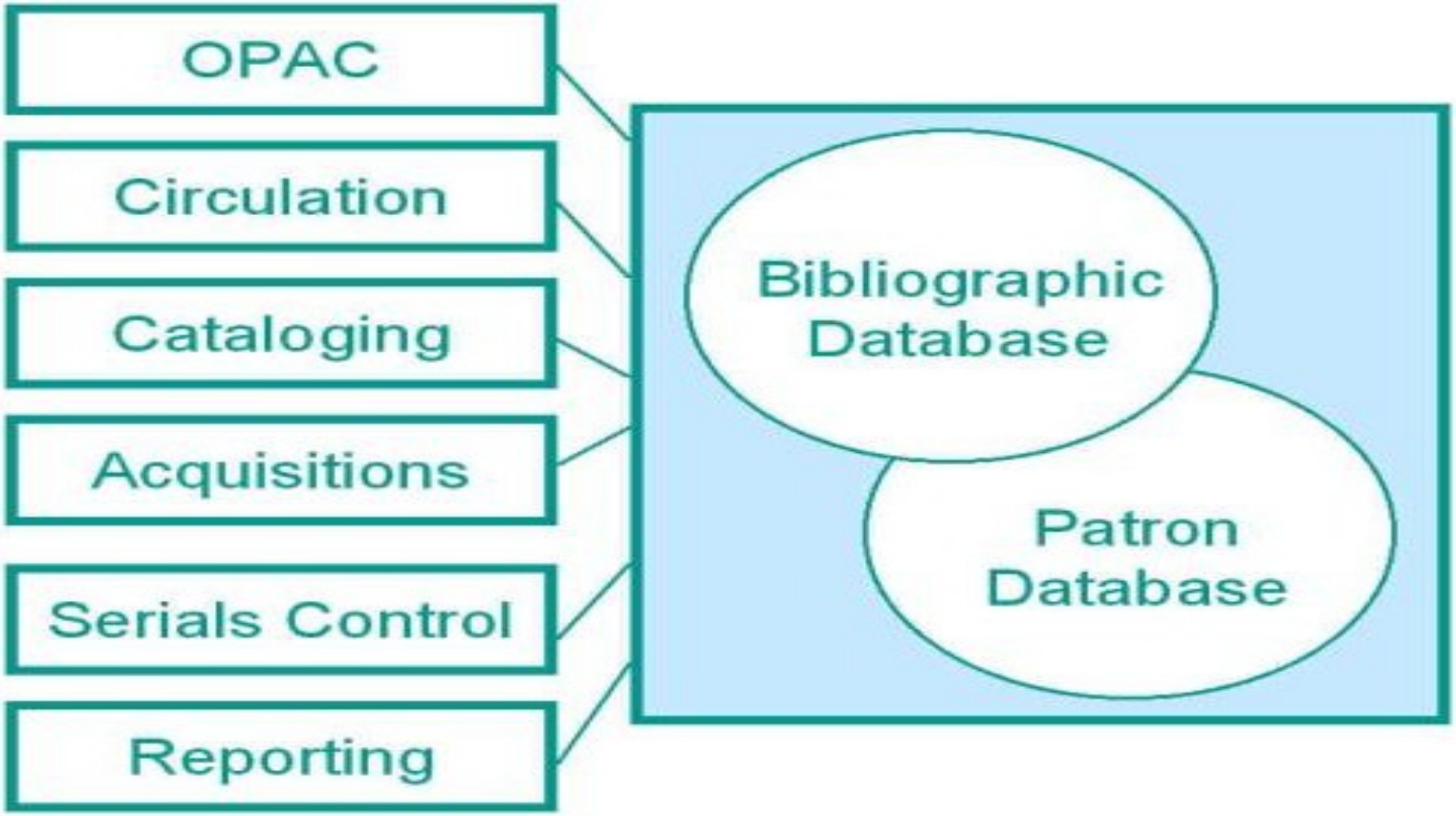
Namen Z39.50

- Urejati vse vidike distribuiranega poizvedovanja
- Sestavni del različnih integriranih sistemov za avtomatizacijo knjižnice (ILS - Integrated Library Systems)

Distribuirano poizvedovanje



Integriran sistem za avomatizacijo knjižnic (ILS)



Funkcionalnosti Z39.50

- Iskalna funkcija
- Avtentikacija
- Kontrola virov
- Funkcija “Explain”
- Brskanje
- Definiranje formata bibliografskih zapisov

+

“Extended services”

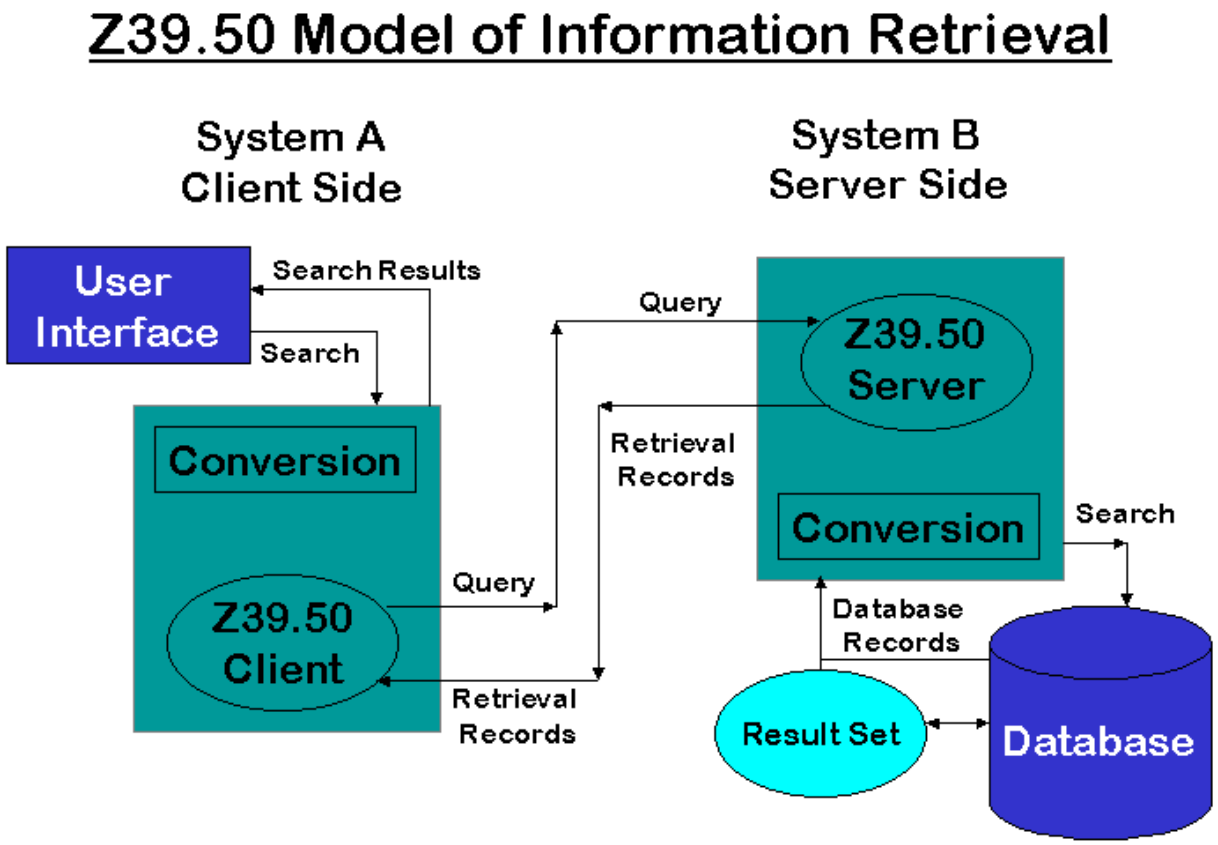
Extended services

- Shranjevanje iskalne zahteve za kasnejšo rabo
- Shranjevanje rezultatov za kasneje
- Definiranje urnika periodičnega poizvedovanja
- Naročilo izvoda
- Posodobitev podatkovne baze
- Ustvaritev izvozne specifikacije

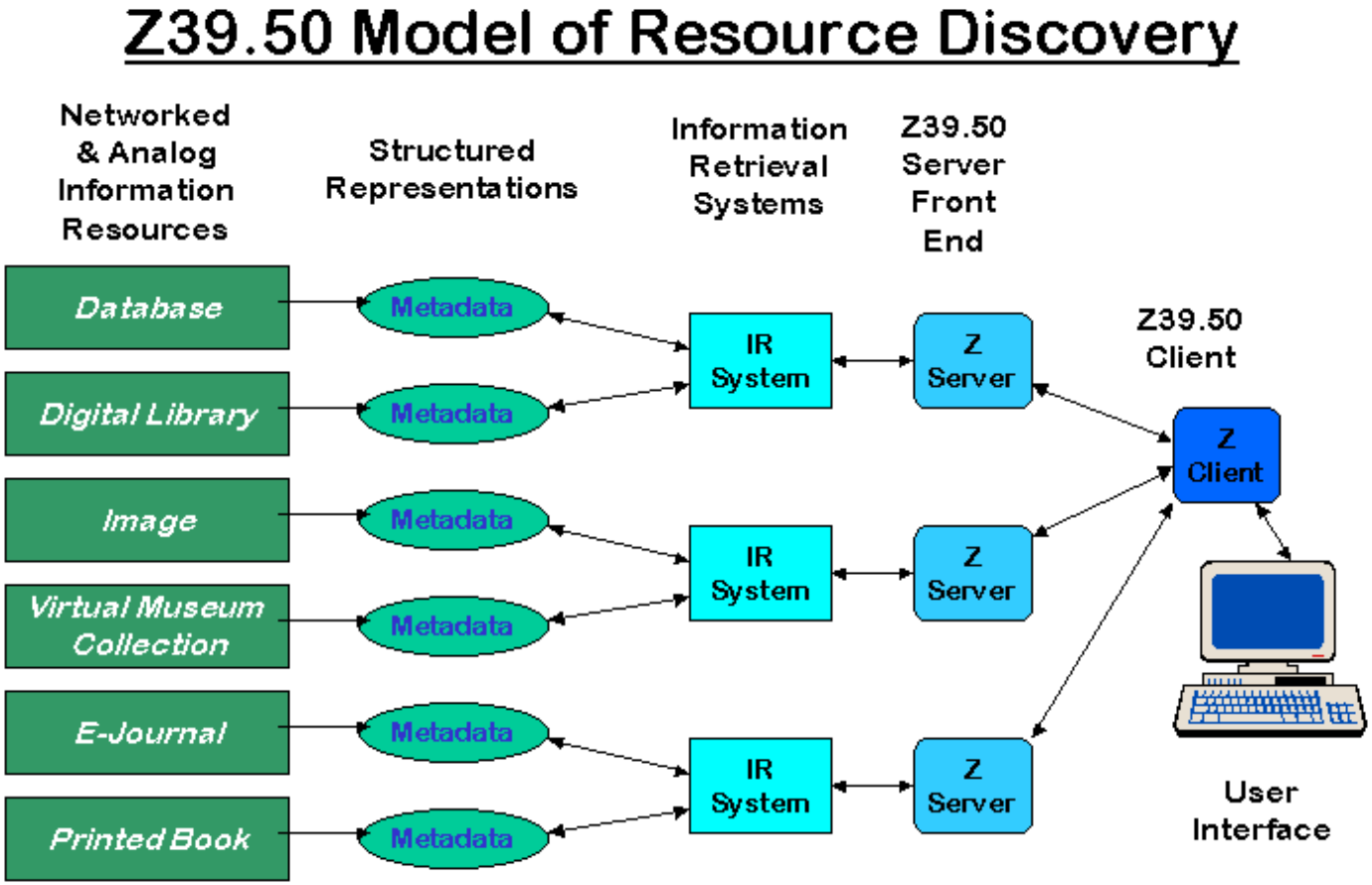
Funkcije Z39.50

- Inicializacija (»Initialisation«)
- Iskanje (»Search«): vsebuje servis oz. podfunkcijo »Search«
- Prevzem (»Retrieval«)
- Izbris seta rezultatov (»Result-set-delete«)
- Avtentikacija oziroma kontrola dostopa (»Access Control«)
- Zaračunavanje / Kontrola resursov (»Accounting / Resource Control«)
- Sortiranje (»Sort«)
- Brskanje (»Browse«)
- Dodatni servisi (»Extended services«)
- Obrazložitev (»Explain«)
- Prekinitev (»Termination«)

Primer delovanja Z39.50



Distribuirano poizvedovanje



Z39.50 poizvedba

- Uveden dodaten nivo abstrakcije
- Sintaksa poizvedbe tipično v Type-1 poizvedbi (RPN)
- Semantični nivo v kakšnem izmed setov atributov: najpogosteje bib-1

http://www.loc.gov/z3950/agency/defns/bib1.html

Z39.50 primer

- 1=21 kamen OR mineral
- @attr 4=1 @and @attr 1=1 "bob dylan" @attr 1=4 "slow train coming“

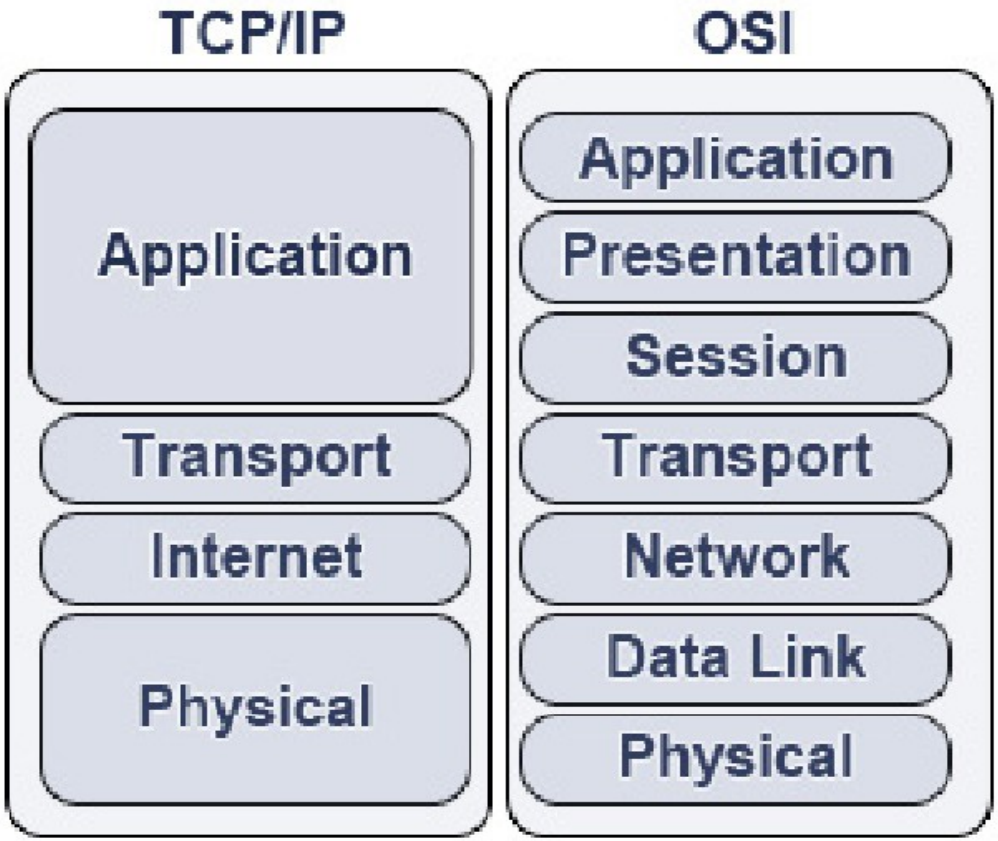
http://www.loc.gov/z3950/agency/defns/bib1.html

Vir primera: http://www.indexdata.com/yaz/doc/tools.html

Splet

- Premik od ILS paradigme k spletni paradigmi
- Spletne tehnologije (W3C) so standardizirane in preizkušene, ter široko uporabljene
- Kot programer se nočemo ukvarjati z zelo nizko nivojskimi stvarmi
- Kot uporabnik si prav tako želimo komodnosti

OSI in TCP/IP modela



Problemi distribuiranega poizvedovanja v spletnem okolju

- Kako izmed množice podatkovnih baz, ki so na voljo, omogočiti uporabniku, da bo lahko izbral tako, ki jo potrebuje.
- Kako paralelno/simulatano poizvedovati po izbranih podatkovnih bazah, ki po možnosti uporabljajo različne oblike poizvedbe oziroma načine formiranja poizvedbe, v spletenem okolju, kjer zaradi okvar, vzdrževanj, namenskega začasnega umika, vsi viri niso vedno dosegljivi.
- Kako združiti rezultate vrnjene s strani različnih iskalnikov, virov, ki so različno formatirani (v

SRU

- Tri funkcije/operacije:
  - »search/retrieve«
  - »explain«
  - »scan«
- Naslanja se na W3C standarde
- XML
- Bazira na HTTP, zato statični protokol
- Različni načini prenosa: via GET, via POST, via SOAP

Search/Retrieve

- Semantični del: kontekstni set (metapodatkovna shema, polja opisa) isti kot na strežniku
- Sintaktični del: CQL
  - Dva tabora poizvedovalnih jezikov: v prvem so zmogljivi, zelo izrazni jeziki, ki pa so težko berljivi in zapisljivi s strani neekspertov (npr. SQL, PQF in XQuery), v drugem taboru pa preprosti in intuitivni jeziki, s katerimi pa zato težko izrazimo zapletenejše koncepte (npr. CCL in googlov jezik za iskanje)
  - CQL kot vmesna verzija obeh tipov jezikov

CQL

- riba
- dc.title any riba or dc.description any riba
- dc.title any/stem "računalništvo avtomatizacija"
- "riba" sortBy dc.title/ignoreCase
- veverica sortby steviloNog/number
- dc.title = l*s
- dc.title = l?s

Primer

http://z3950.loc.gov:7090/voyager?version=1.1&que

Ostali dve funkciji

- Explain

  http://z3950.loc.gov:7090/voyager?version=1.1&operation=explain

- Scan

  Isto kot pri Z39.50: zahteva seznam mogočih

  terminov znotraj seznama indeksiranih terminov

Primerjava Z39.50 in SRU - uvod

- Z39.50:
  - S pomočjo Z39.50 je moč izvesti 9 operacij: operacija kot zahteva skupaj z ustreznim odgovor, vključujoč izmenjana sporočila
  - 11 funkcij pridobivanja informacij, vsaka sestoji vsaj iz ene ali več storitev
- SRU:
  - zgolj tri funkcije: »Search/Retrieve«, »Explain« ter »Scan«

Primerjava Z39.50 in SRU - okolje

- Z39.50:
  - predspletni protokol
  - ILS paradigma
- SRU:
  - za delovanje se naslanja na W3C standarde (tako kot OAI-PMH)
  - paradigma spletnih storitev (“Web services”)

Primerjava Z39.50 in SRU - poizvedba

- Z39.50:
  - Tipično Type-1 oblika poizvedbe z bib-1 setom atributov
- SRU:
  - CQL
  - Kontekstni set isti kot v podatkovni bazi sami (ni nivoja semantične abstrakcije)

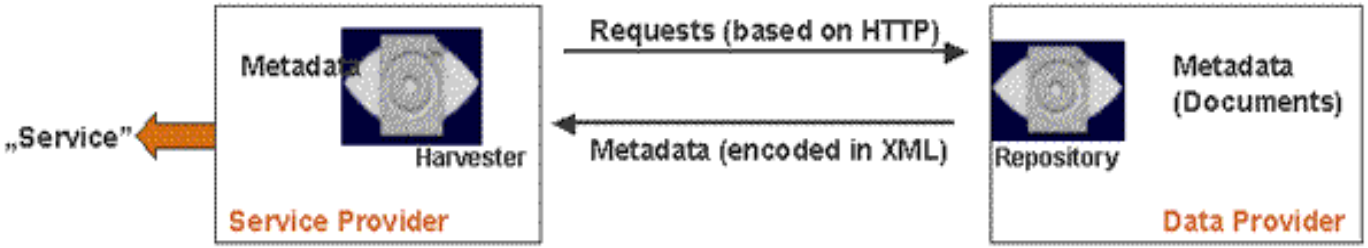
OAI-PMH

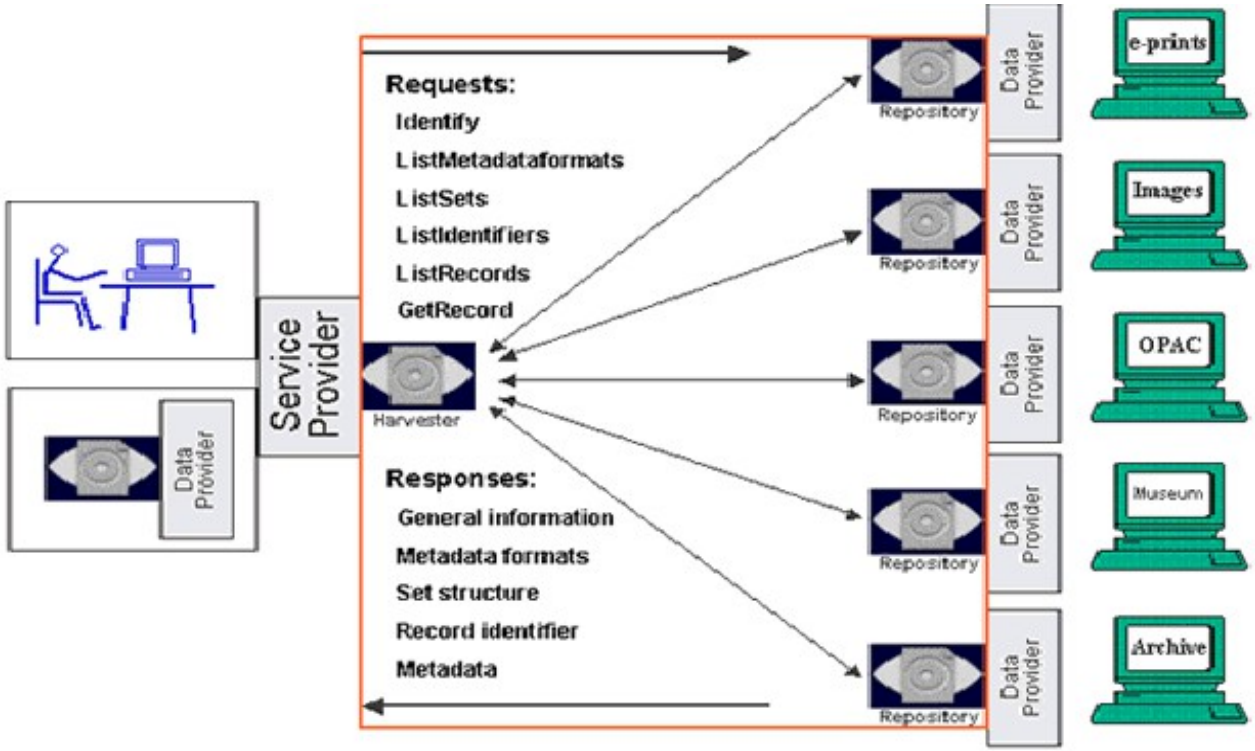
- Primarni namen OAI-PMH je definirati standarden način, kako prenesti metapodatke iz točke A v točko B. Posredno pa je namen **omogočit širjenje in zbiranje (agregacijo) metapodatkov**, ki opisujejo uporabne informacijske vire.
- Tehnologije:
  - splet
  - Dublin Core
  - XML

OAI-PMH terminologija

- OAI-PMH **ponudniki podatkov** (data provider) posedujejo zbirko primarnih dokumentov (običajno) in metapodatkov, ki opisujejo to vsebino (vedno)
- Ponudnik podatkov (data provider) dajejo te metapodatkovne zapise na voljo po pravilih, ki jih določa protokol
- **Ponudniki storitev** (service provider) spet v skladu s protokolom žanjejo (harvest) s strani ponudnikov podatkov
- Vir

Delovanje OAI-PMH





Vir: http://www.oaforum.org/tutorial/english/intro.htm

OAI-PMH napogostejši način rabe

- »one-stop shopping« (mesto, ki nudi poizvedbo po več virih hkrati) model informacijskega poizvedovanja.
- OAI-PMH ponudnik storitev požanje metapodatke iz različnih, ponavadi široko distribuiranih ponudnikov podatkov, združi požete metapodatke v neko podatkovno strukturo ali lokalno podatkovno bazo in nato uporabniku omogoči hkratno iskanje po vseh teh virih s pomočjo enotnega vmesnika
- Uporabniki si prihranijo napor obiska vsakega posameznega ponudnika podatkov.

Obveznosti implementatorjev

- DC kot eden izmed metapodatkovnih formatov
- Vsi formati v skladu z javno XML shemo
- Vsak metapodatkovni objekt v OAI-PMH repozitoriju ima edinstven in trajen OAI identifikator

- Poljubna implementacija setov, politike izbrisa

6 glagolov + argumenti

- »Identify«
- »ListSets«
- »ListMetadataFormats«
- »ListIdentifiers«
- »ListRecords«
- »GetRecord«

Primer OAI-PMH poizvedb (glagoli + argumenti)

- http://eprints.fri.uni-lj.si/cgi/oai2?verb=ListRecords&metadataPrefix=oai_dc
- http://eprints.fri.uni-lj.si/cgi/oai2?verb=GetRecord&identifier=oai:generic.eprints.org:4&metadataPrefix=oai_dc
- http://eprints.fri.uni-lj.si/cgi/oai2?verb=ListRecords&from=2002-05-01T14:15:00Z&until=2011-05-01T14:20:00Z&metadataPrefix=oai_dc

Primerjava Z39.50/SRU in OAI-PMH (Cole in Foulonneau, 2007)

- Lokacija gradiva samega (gradiva, na katerega bibliografski zapisi kažejo)
- Nadzor nad gradivom
- Lokacija metapodatkovnih zapisov
- Semantična interpretacija in sama uporaba iskalne zahteve
- Omejenost hitrosti izvedbe iskalne zahteve
- Zastarelost metapodatkovnih zapisov
- Normalizacija pred iskanjem (ponudnikov storitev)
- Integracija iskalnih rezultatov (sortiranje in spojitev)

- Različna izvora

Zaključki

- Z39.50 bržkone zastarel
- SRU ali OAI-PMH? – odvisno od potrebe, najraje oba
- Kako izvesti iskanje: SRU in OAI-PMH sta ubrala dva različna pristopa

Primeri za spletno okolje

- http://prevoz.org