Kako arhivirati Web

Miroslav Milinović
Sveučilišni računski centar Sveučilišta u Zagrebu
<miro@srce.hr>

Mirna Willer
Nacionalna i sveučilišna knjižnica, Zagreb
<mwiller@nsk.hr>

50. kolokvij knjižnice Instituta "Ruđer Bošković"

Zagreb, 1. prosinca 2004.

Sadržaj

- Informacijski prostor Weba
- Iskustva s hrvatskim prostorom Weba
- O arhiviranju Weba
- Zaključak (što nam je činiti?)

Web: informacijski servis

- jednostavno publiciranje
  - manje barijera / lakši pristup
  - brzo i efikasno publiciranje (posebno za dinamičke izvore informacija)
- informacije su distribuirane
- upravljanje informacijskim prostorom je teško, mogućnosti su ograničene
- novi odnosi između autora, izdavača, distributera, posrednika i korisnika (potrošača)

Informacijski prostor Weba

- “površinski” (*publicly indexable*) Web
  - veljača 1999., *Lawrence and Giles, NEC Institute*
    - 800 milijuna stranica, 15 (6) TB informacija
    - sadržaj: 83% com, 6% sci/edu, 1.5% porn
    - 60% Weba je indeksirano / katalogizirano
  - siječanj 2000., *Inktomi & NEC Institute*
    - više od 1 milijarde Web stranica
    - top-level domene: 55% .com, 8% .net, 4% .org, 1% .gov
  - u stalnom rastu:
    - ≈ 41 milijun Web sjedišta (lipanj 2003., Netcraft)
    - ≈ 5 milijardi Web stranica (2003.)

Informacijski prostor Weba

- 85% korisnika rabi pretraživačke mahanizme ili tematske kataloge kako bi pronašli informacije

  Steve Lawrence, Lee Giles , Nec Institute, veljača 1999.

- korisnici smatraju da je Internet važan izvor informacija
  - 2/3 korisnika smatra da je Internet važan ili vrlo važan izvor informacija
  - 53%(47%) smatra TV (radio) jednako važnim

    Center for Communication Policy, UCLA, kolovoz 2000.

- visoka očekivanja korisnika:
  - 97% korisnika očekuje pronači traženo

    Pew Internet & American Life, 2002.

Informacijski prostor Weba

- problemi inačica i duplikata:
  - caching, mirroring
- 40% od 800 milijuna stranica su duplikati

  FAST, 2000.
- 30% Web stanica su kopije

  Shivakumar and Garcia-Molina, 1998.
- “Deep” Web
  - 400 do 550 puta veći od “surface” Weba
  - 7500 TB podataka

    The Deep Web: Surfacing Hidden Value; BrightPlanet.com, srpanj 2000.

Problemi?

- velika očekivanja korisnika
- alati i mehanizmi za upravljanje informacijskim prostorom Weba u stalnom razvoju (nikad dovoljno dobri)
- informacijski prostor nije (dobro) organiziran
- nepouzdana (nesigurna):
  - kvaliteta informacija
  - integritet informacija
  - povjerenje u izvor informacija (autentičnost)
- problem identifikacije resursa (URL)

Iskustava Srca s hrvatskim prostorom Weba

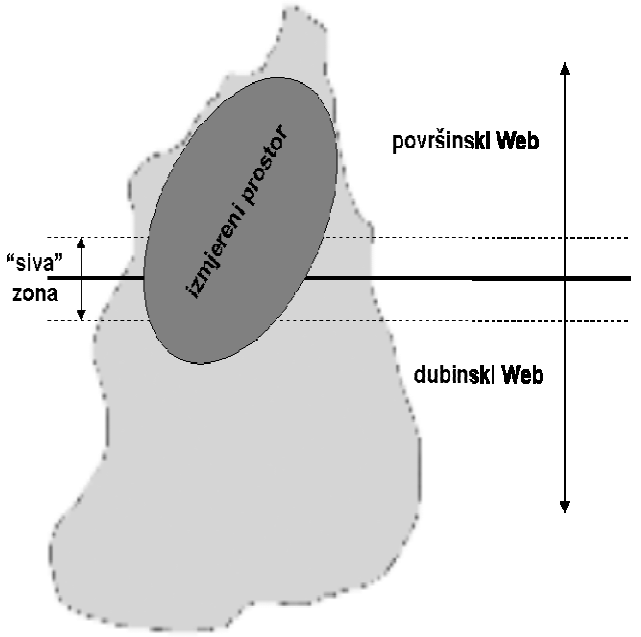
- Mjerenje hrvatskog Web prostora (**MWP**)
  - http://www.srce.hr/mwp/
  - Kontakt adresa mwp@srce.hr
  - započeto 2002. godine
- Projekt uspostave sustava za preuzimanje i arhiviranje obveznog primjerka hrvatskih mrežnih publikacija (**DAMP**)
  - zajednički projekt NSK i Srca
  - započet u studenom 2003. godine
  - cilj: uspostaviti sustav za preuzimanje i arhiviranje uz očuvanje autentičnosti sadržaja, oblika i funkcionalnosti mrežnih publikacija
- Projekt uspostave sustava za preuzimanje i arhiviranje mrežnih resursa za potrebe HIDRA-e (**AMD**)
  - zajednički projekt HIDRA-e i Srca
  - započet u svibnju 2004. godine
  - cilj: uspostaviti sustav za preuzimanje i arhiviranje mrežom dostupnih dokumenata s odabranog skupa Web sjedišta

MWP: što se i kako mjerilo?

- Provedena mjerenja:
  - MWP1: 29.03.-07.05.2002.
  - MWP2: 14.05.-22.07.2003.
  - MWP3: 08.09.-25.11.2003.
- Predmet mjerenja:
  - elektronički resursi dostupni HTTP / HTTPS protokolom s poslužitelja u .hr vršnoj internetskoj domeni
- Mjerilo se:
  - veličinu
  - korištene formate zapisa (prema MIME standardu)
  - obim i sadržaj metapodataka

Rezultati mjerenja

- MWP1 vs. MWP3:
  - MWP1 (2002.)
    - “HTTP resursi”
    - ≈ 320 GB (389 GB)
    - 5.145.383 obrađena resursa
  - MWP3 (2003.)
    - “HTTP/HTTPS resursi”
    - ≈ 548 GB
    - 7.125.879 obrađenih resursa (6.433.902 uspješno)
    - text/html: 3.194.548 resursa, prosječne veličine 28.902 byta



MWP vs. svijet

- Web je “*small and simple*”:
  - MWP-1(2002.):
    - 320 GB i ≈ 6 milijuna resursa
    - 5 MIME tipova čini više od 90% obrađenih resursa
  - švedski Web (Hakala, 1999.):
    - 300 GB i ≈ 7,5 milijuna resursa
    - 4 MIME tipa pokrivaju 97% resursa
- Uporaba metapodataka:
  - Lawrence & Giles (1999.):
    - 34% Web resursa ima META oznaku / 0,3% rabi DC
    - 123 različita oblika META oznake
  - MWP:
    - MWP-1: 31% Web resursa ima META oznaku / 0,09% rabi DC
    - MWP-2: 43,1% Web resursa ima META oznaku / 2,31% rabi DC
    - broj različitih vrijednosti NAME atributa META oznake:
      MWP1: 743 / MWP2: 645 / MWP3: 666

MWP: zaključak

- rezultati odgovaraju očekivanjima i sličnim istraživanjima provedenim u svijetu
- (površinski) Web je i dalje jednostavan: rabimo mali broj različitih formata
- autori ne brinu dovoljno o metapodacima
- dinamički web, inventivne, ali i nestandardne uporabe Web tehnologija čine mjerenje sve složenijim

Projekt DAMP

- Projekt uspostave sustava za preuzimanje i arhiviranje obveznog primjerka hrvatskih mrežnih publikacija
  - zajednički projekt NSK i Srca, uz suradnju tvrtke UNIBIS
  - započet u studenom 2003. godine
  - **Cilj:** uspostava sustava za preuzimanje i arhiviranje obveznog primjerka hrvatskih publikacija na internetu uz očuvanje, u najvećoj mogućoj mjeri autentičnosti njihova sadržaja, oblika i funkcionalnosti, a u svrhu njihove dugoročne zaštite i korištenja u budućnosti. Cilj projekta usklađen je s temeljnom zadaćom NSK da prikuplja, obrađuje, čuva i daje na korištenje hrvatske publikacije na svim medijima.

- DAMP (**D**igitalni **A**rhiv **M**režnih **P**ublikacija) ver. 1.0.
  - modularan, proširiv, jednostavna uporaba za korisnika/arhivara
  - utemeljen na Open Souce programskoj podršci
  - razvijen na temelju iskustava Srca u području istraživanja Weba (projekt MWP)

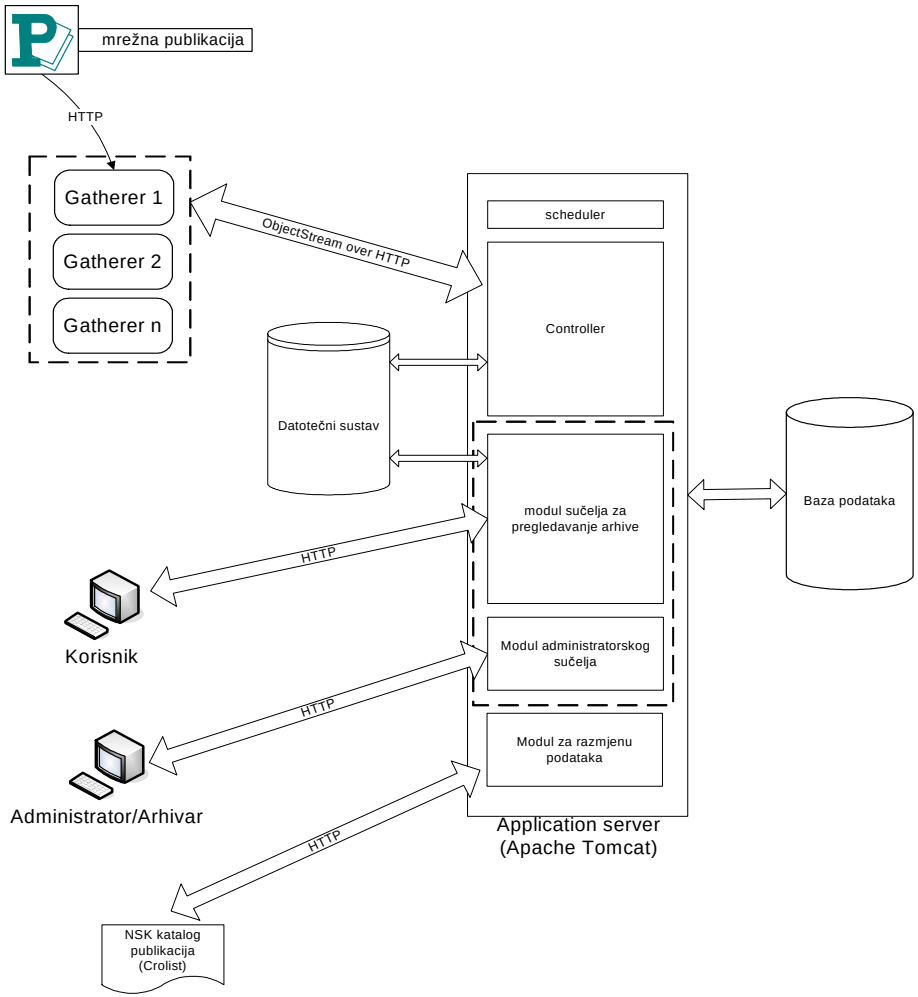
DAMP: temeljni principi

- Pobiru se i arhiviraju samo one publikacije koje su bibliografski identificirane i obrađene, tj. poslovanje mrežnim publikacijama podliježe svim postupcima knjižničnog poslovanja kao s građom koja nije mrežna.
- Sustav DAMP kao izvor osnovnih informacija o publikacijama (Web sjedištima, kandidatima za pobiranje) rabi katalog NSK (CROLIST)

Dijelovi sustava DAMP

- **Gatherer**
  - podsustav za pobiranje Web sjedišta (publikacije)
- **Controller**
  - podsustav za kontrolu pobiranja i arhiviranje rezultata pobiranja
- **Scheduler**
  - podususstava za raspoređivanje pobiranja
- **Skladište podataka**
  - čine ga baza podataka i posebno organizirani datotečni sustav
- **Web sučelje** za pristup arhivu i upravljanje sustavom
- **Modul za razmjenu podataka** s Katalogom NSK

Funkcionalni model sustava DAMP



DAMP: upravljanje pobiranjem

- administrator/arhivar mora obraditi svaku novu publikaciju:
  - pregledati konkretno Web sjedište
  - odrediti frekvenciju (vrijeme i učestalost) pobiranja
  - odrediti parametre pobiranja (prije svega **dubinu**)
  - provjeriti ishod prvog (probnog pobiranja) i eventualno dodatno podesiti uvjete/parametre
- temeljem postavljenih uvjeta i parametara scheduler uvrštava konkretno pobiranje u red
- gatherer(i) izvršava(ju) pojedina pobiranja prema tom redu (pod nadzorom controllera)

DAMP: pristup arhivu

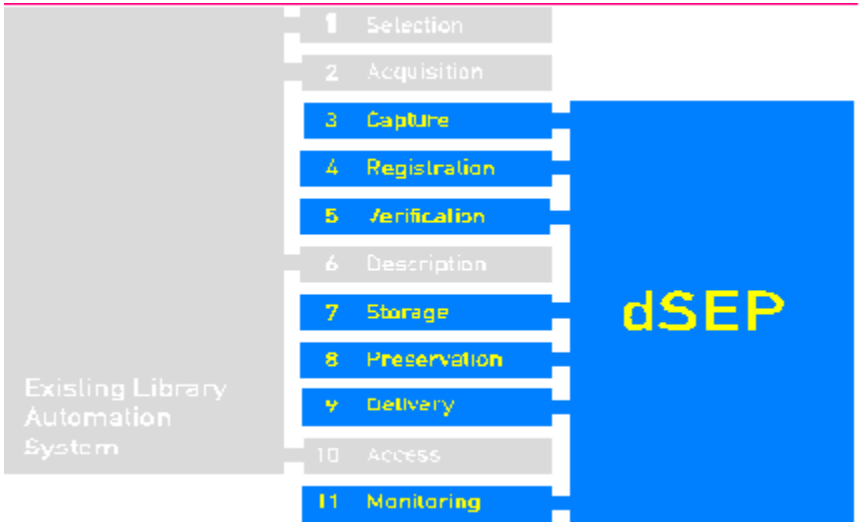
- putem Web sučelja
- za svako pobiranje pojedine publikacije omogućuje:
  - pregledavanje arhivskog primjerka
  - pregledavanje podataka o pobiranju
    - osnovni podaci o pobiranju (veličina, status, ...)
    - distribucija tipova prikupljenih datoteka
    - metapodaci
  - pregledavanje zapisa o pobiranju (log)
  - izravni pristup direktoriju s rezultatom pobiranja

Koncept i funkcionalni model integracije sustava DAMP i CROLIST

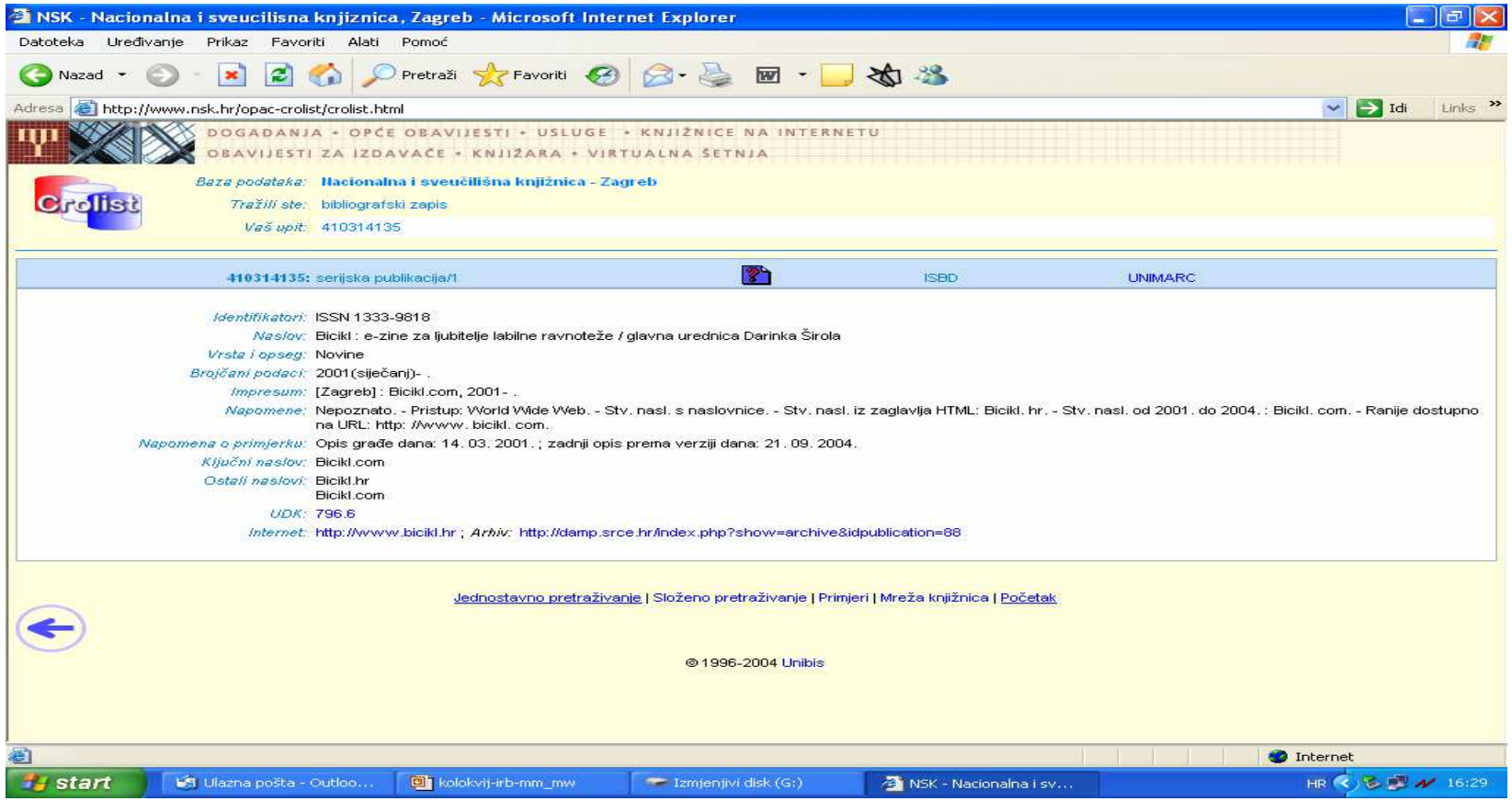
- Koncept: Pobiru i arhiviraju samo one publikacije koje su bibliografski identificirane i obrađene, tj. poslovanje mrežnim publikacijama podliježe svim postupcima knjižničnog poslovanja kao s građom koja nije mrežna, od njene identifikacije do pronalaženja u WebPAC-u NSK.
- Posebnosti poslovanja s mrežnom građom odnose se na upravljanje njihovim arhiviranjem u poslovnom procesu interakcije sustava DAMP i CROLIST.

Integriranje digitalnog arhiva s LIS-om (Projekt NEDLIB)

1. Odabir za izgradnju fonda
2. Nabava
3. Pobiranje
4. Registracija
5. Provjera
6. Opis/Katalogizacija
7. Pohrana/rukovanje
8. Zaštita
9. Isporuka
10. Pristup
11. Nadzor



Bicikl : e-zine u WebPAC-u



DAMP: izazovi

- dinamički web, inventivne, ali i nestandardne uporabe Web tehnologija čine arhiviranje složenim (nemogućim?)
- dubinski web - nepoberiv (?)
- publikacije zaštićene nekim mehanizmom autentikacije/autorizacije
- off-line arhive nakladnika
- uvjeti / ograničenja na javni pristup arhivu
- DAMP-dijagnostika u funkciji procjene kvalitete arhivskog primjerka
- minimalni uvjeti za publikacije (upute za nakladnike)
- osigurati suradnju nakladnika

Arhiviranje Weba

- izazovno, ali globalno/sveobuhvatno (ne)moguće (?)
- nužna/važna selekcija/izbor resursa koji se pobiru
- mogući opći kriteriji:
  - domena
  - jezik
  - sadržaj
  - vrsta publikacija
  - oblik/način objavljivanja
  - pristup
  - format
  - ...
- metode pobiranja:
  - dostava, pobiranje, pobiranje uz prethodnu selekciju, ...

Projekti arhiviranja Web u svijetu

Zemlja	Projekt	Metoda	Pristup	Veličina
Australija	PANDORA (NLA)	Odabir	Da	353 GB
Austrija	AOLA	Pobiranje	Ne	448 GB
Finska	Sveučilišna Knjižnica Helsinki	Pobiranje	Ne	401 Gb
FR	BNF	Odabir i pobiranje	Ne	< 1 TB
Švedska	Kulturarw	Pobiranje	Ograničen pristup	4.5 TB
VB	Britain on the Web, BL	Odabir	Ne	30 MB
SAD	Internet Archive	Pobiranje	Da	> 150 TB
SAD	MINERVA, LC	Odabir	Ne	35 sites
HR	NSK	Odabir	?	ca 800 publ.

Klarin, S. AKM8: Arhiviranje sadržaja weba u projektima nacionalnih knjižnica i projektu "Internet Archive"
Day, Michael. Collecting and preserving the World Wide Web : version 1.0 – 25 february 2003. Str. 18.

Što nam je činiti? (kao autorima/nakladnicima)

- upoznajmo i poštujmo standarde i preporuke
  - meta podaci
  - Standards for Robots Exclusion (REP, ROBOTS META oznaka)
  - W3C WAI
  - ...
  - http://useit.com/
  - domaće incijative i projekti
  - ...
- Izbjegavajmo inventivnu, ali nestandardnu (neuobičajenu) uporaba Web tehnologija
  - Uporaba tehnologije radi tehnologije same
  - "best viewed with any browser"

Kako arhivirati?

- Definirati uvjete / kriterije
- Definirati metode / tehnike
- Educirati nakladnike / autore, ali i korisnike
- Pratiti razvoj (Web) tehnologija

- Nema kvalitetnog arhiva bez nadzora arhivara!

Sadržaj

- Informacijski prostor Weba
- Iskustva s hrvatskim prostorom Weba
- O arhiviranju Weba
- Zaključak (što nam je činiti?)

Kako arhivirati Web

Miroslav Milinović
Sveučilišni računski centar Sveučilišta u Zagrebu
<miro@srce.hr>

Mirna Willer
Nacionalna i sveučilišna knjižnica, Zagreb
<mwiller@nsk.hr>

50. kolokvij knjižnice Instituta "Ruđer Bošković"

Zagreb, 1. prosinca 2004.