Tehnološki aspekti arhiviranja weba u Hrvatskoj

Miroslav Milinović
Sveučilište u Zagrebu, Sveučilišni računski centar
<Miroslav.Milinovic@srce.hr>

Stručni skup povodom 10. obljetnice Hrvatskog arhiva weba
Zagreb, 26. rujna 2014.

srce
Sveučilište u Zagrebu
Sveučilišni računski centar

Web – kako je počelo?

Information Management: A Proposal

Tim Berners-Lee, CERN, March 1989, May 1990

"The Web will become a
robust, scalable, adaptive infrastructure,
framework for computation of knowledge,
communication medium."

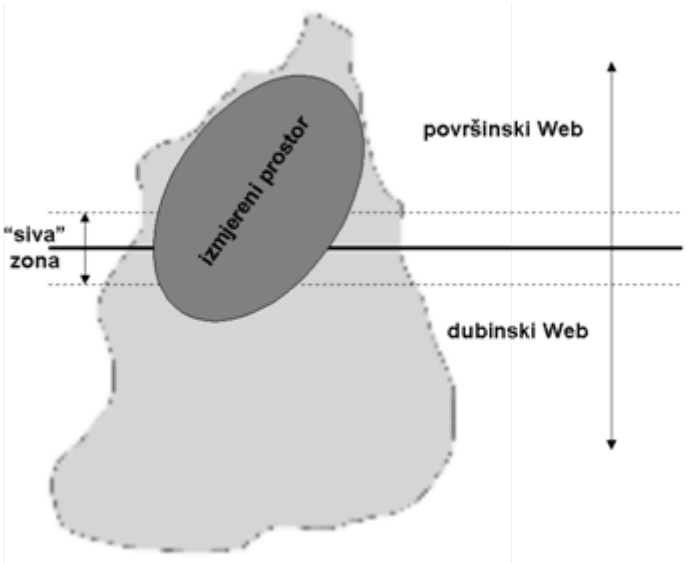
Informacijski prostor Weba

- obiman, složen
- distribuiran, dinamičan
- kontinuirano raste i mijenja se
  - jednostavna upotreba
  - lagano publiciranje
- tehnologiju odlikuje:
  - kontinuirani razvoj
  - raznolika rješenja
  - mnoštvo standarda

Informacijski prostor Weba: izazovi

- kvaliteta informacija
- integritet informacija
- povjerenje u izvor informacija (autentičnost)
- identifikacija resursa (URL)
- pitanje inačica i duplikata (*caching*, *mirroring*)
- može li se Web efikasno:
  - kontrolirati
  - mjeriti
  - pretraživati
  - arhivirati
  - …?

Koliko je Web velik?



dubinski (*deep*) vs. površinski (*indexable*) Web

Harvestiranje (pobiranje) Weba

- harvestiranje = pobiranje
- svrha harvestiranja:
  - indeksiranje (za potrebe tražilica)
  - **arhiviranje (cilj: sačuvati izvorni izgled)**
  - istraživanje (npr. mjerenje ili provjera usklađenosti s normama)
- harvester (*crawler, gatherer, robot)* – program koji obilazi web
  - započinje početnim popisom URL adresa (*seed*)
  - iterativni postupak:
    - dohvat web stranice odnosno svih resursa koji joj pripadaju
    - obrada i pronalaženje poveznica (linkovi na stranice, slike, video, skripte, dokumente, css,…)
    - dodavanje novootkrivenih URL-ova na popis

Iskustva Srca

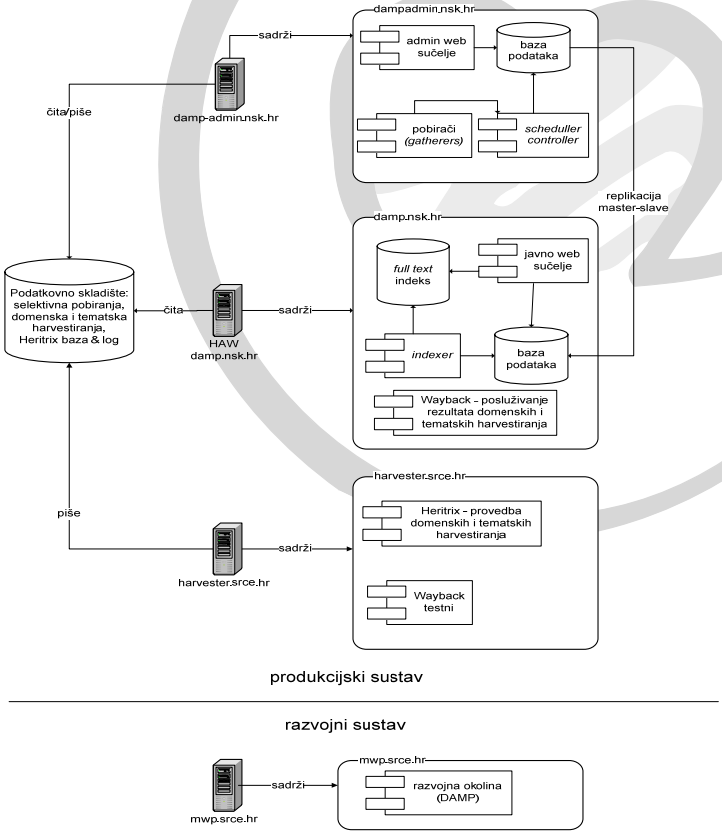
- projekt mjerenja Web-prostora (MWP)
  - od 2002. do 2008. godine
  - http://www.srce.hr/mwp
- suradnja s NSK-om na razvoju i održavanju Hrvatskog arhiva Weba
  - kontinuirano od 2003. godine
  - http://haw.nsk.hr
  - programski sustav DAMP
- suradnja s Digitalnim informacijsko-dokumentacijskim uredom Vlade RH
  - kontinuirano od 2004. godine
  - sustav AMD (arhiv mrežnih dokumenata) koji je temelj sustava DAMIR (Digitalni arhiv mrežnih izvora Republike Hrvatske)

Hrvatski arhiv Weba (HAW)

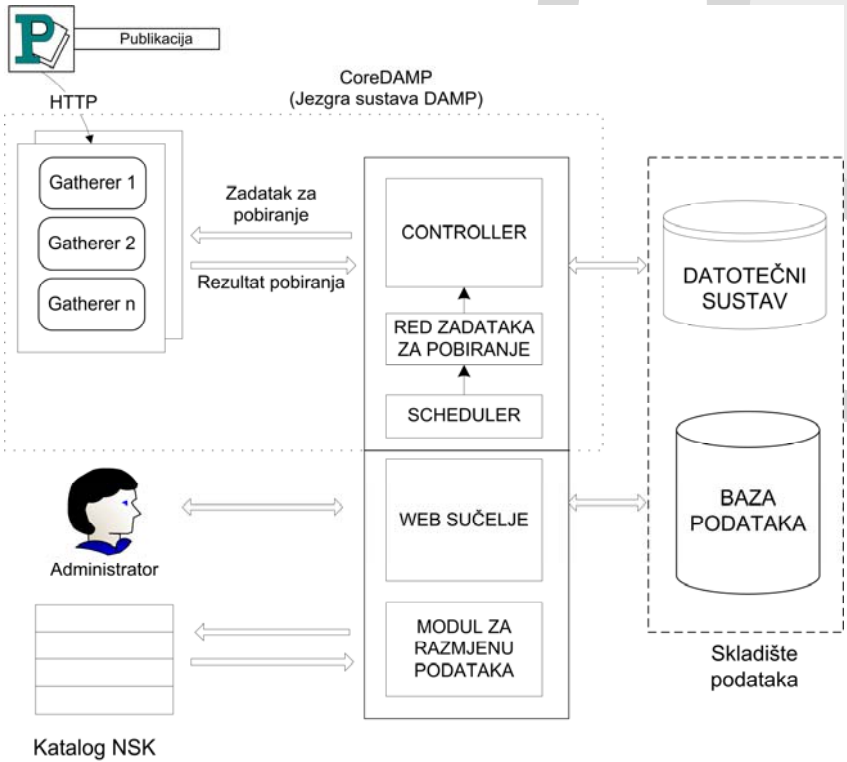
- dva temeljna dijela:
  - **sustav za selektivno pobiranje/arhiviranje**
    - tehnička pozadina je sustav **DAMP** razvijen u Srcu
      - modularan, proširiv i prilagodljiv
      - utemeljen na otvorenom kodu
      - u produkciji od 2004. godine
  - **sustav za arhiviranje nacionalne domene i tematska arhiviranja**
    - modificirana programska podrška: **Heritrix / Wayback**
    - prvo arhiviranje .hr domene provedeno je u vremenu od 18. srpnja do 18. kolovoza 2011.
      (ukupno je pobrano više od 56 milijuna datoteka ukupne veličine od preko 3.1 TB)

Arhitektura HAW-a

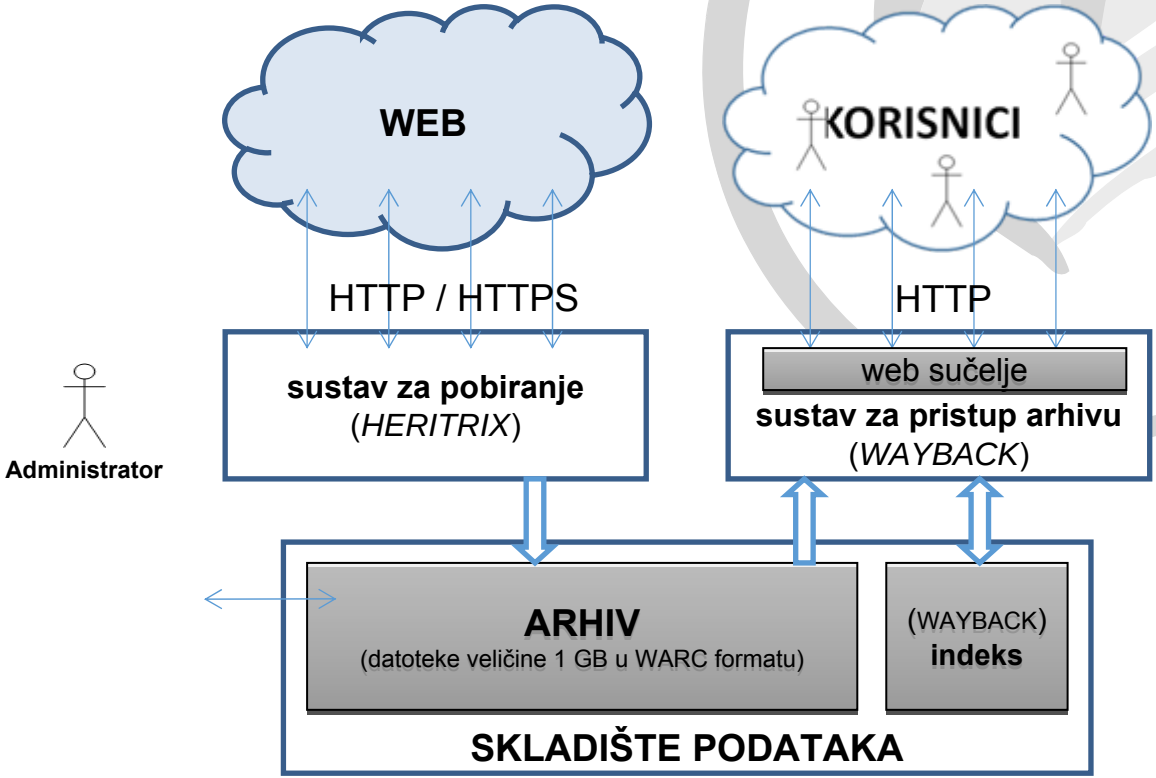
- **HAW 2014**
  - 31 TB
  - 4 poslužitelja
  - OpenSource / LAMP
  - udomljen u Srcu



HAW: Funkcionalni model sustava DAMP



HAW: Arhitektura sustava za arhiviranje nacionalne domene i tematska arhiviranja



DAMP vs. Heritrix

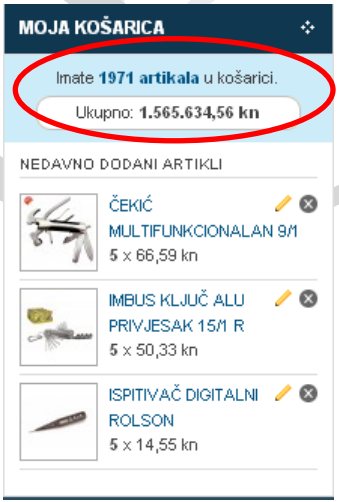
Heritrix	DAMP
Za prikaz arhiviranih sadržaja potrebna je dodatna programska podrška: Wayback	Prilikom harvestiranja radi se prilagodba za prikaz pa nije potrebna dodatna programska podrška
Rezultati: arc / warc (potrebno manje diskovnog prostora i lakše održavanje)	Rezultati: mirror (1 sjedište = 1 direktorij); velik broj datoteka
Fleksibilnost, velik broj opcija za fino podešavanje	Jednostavnost korištenja
Sprema zaglavlja HTTP upita i odgovora	Skalabilnost (posao se lako raspodijeli na dodatne poslužitelje)

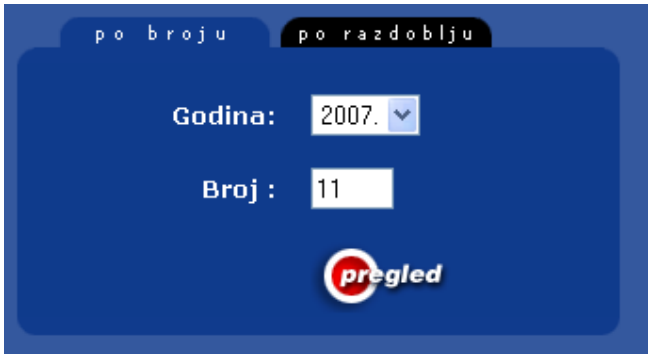
Izazovi pri arhiviranju: planiranje

- izbor resursa koji se arhiviraju
  - protokoli (HTTP / HTTPS),
  - izbor internetske domene
  - početni popis adresa (*seed*)
  - dubina, broj resursa, veličina po sjedištu
  - kriterij završavanja
- umetnuti (*embedded*) resursi
  - slike, video, skripte (Facebook umetci, Google Ads,…), elementi dizajna tj. stilovi, *frame*-ovi…
- preusmjeravanja
  - http://www.nesto.hr/ -> http://nesto.com/
- pridržavanje robots.txt pravila

Izazovi pri arhiviranju: provedba

- konfiguracije CMS alata - robots.txt datoteke koje zabranjuju pristup nekim / svim sadržajima
- „beskonačna" web sjedišta - isti sadržaj na svakoj stranici ima drugačiji link
- Web katalozi / prodaja
- forumi – uz svaki odgovor standardni linkovi
- galerije – uz svaku sliku link na ocjeni sliku
- pogrešna detekcija linkova koji to nisu (javascript, …)





Zaključak

- pobiranje / arhiviranje je moguće
  - uz određene ograde
  - moguće je izraditi *snap-shot* (sliku) „nepoberivih” sjedišta
- (površinski) Web je i dalje jednostavan
  - rabimo mali broj različitih formata
- autori ne brinu dovoljno o standardima i mogućnosti arhiviranja
- dinamički web, inventivne, ali i nestandardne uporabe Web tehnologija čine pobiranje sve složenijim
- izazovi:
  - definiranje opsega pobiranja
  - uobičajeni izazovi za tražilice (forumi, interaktivni katalozi, ...)
  - uloženi resursi (oglasi, društvene mreže, ...)
  - nove web-tehnologije i njihova primjena

damp@srce.hr

Miroslav.Milinovic@srce.hr