

UPORABA SVETOVNEGA SPLETA KOT VIRA INFORMACIJ V KNJIŽNICI

Miro Pušnik, Centralna tehniška knjižnica Univerze v Ljubljani

UDK 681.324WWW

Povzetek

Obsežnost svetovnega spleta (WWW) in vsebinsko nestandardizirana izdelava HTML dokumentov ovirata uporabo svetovnega spleta kot vira uporabnih informacij. Toda če poznamo organizacijo svetovnega spleta, če pravilno izberemo iskalno strategijo in če znamo oceniti ustreznost dobljenih rezultatov, je lahko svetovni splet koristen vir informacij tudi v knjižnici. Delo si olajšamo tako, da si izdelamo zbirko uporabnih virov informacij na svetovnem spletu.

UDC 681.324WWW

Summary

The use of the World Wide Web (WWW) as a source of relevant information in a library is hindered by its extensiveness and unstandardized contents of HTML documents. However, on condition that we are acquainted with the organization of the WWW, if we choose the right searching methods, and if we know how to evaluate search results the WWW can be quite a useful source in a library, too. Our work can be facilitated by forming a collection of usefull WWW information sources.

1. Uvod

Svetovni splet (WWW, World Wide Web, W3...) je obsežna in neurejena množica informacij v obliki dokumentov HTML (Hypertext Markup Language), ki se iz dneva v dan oziroma iz ure v uro veča. Hitremu porastu Svetovnega spleta botruje več dejstev, med drugimi tudi to, da je dokument HTML relativno enostavno sestaviti in predstaviti vsemu svetu (in morda še kam). Na prvi pogled se slednje zdi dobro, če pa vzamemo Svetovni splet kot vir informacij, lahko ob premisleku ugotovimo, da je ravno enostavnost izdelave in postavitve HTML dokumenta na strežnik ena izmed največjih pomanjkljivosti organizacije Svetovnega spleta. Zgodi se, da lahko vsak objavi vse, zato moramo znati vrednotiti elektronske vire, tako kot vrednotimo tudi tiskane.

V CTK (Centralna tehniška knjižnica) smo soočeni z vse večjimi potrebami uporabnikov po informacijah s svetovnega spleta. Uporabniki imajo v knjižnici na voljo dostop do Interneta in velikokrat se zgodi, da prosijo informatorja za nasvet pri iskanju vira informacij. Marsikatera želja je taka, da informatorji sami uporabimo svetovni splet in druge aplikacije Interneta (Usenet, E-pošta, FTP...), da lahko zadostimo potrebam obiskovalcev, vse več pa je tudi informacijskih servisov, ki uporabljajo HTML standard (COBISS, OCLC, DIALOG...).

V tem prispevku bi se rad dotaknil nekaterih slabih točk v organizaciji svetovnega spleta, ki lahko zmanjšajo njegovo uporabnost, če jih ne poznamo, in približal nekatere načine iskanja informacij na svetovnem spletu ter kriterije za vrednotenje virov teh informacij. Tako bomo lahko z veliko verjetnostjo trdili, da je informacija, ki smo jo dobili na svetovnem spletu, verodostojna in kvalitetna.

2. Nekateri značilnosti organizacije svetovnega spleta

Program World Wide Web oziroma svetovni splet je razvil Tim Berners-Lee iz inštituta za fiziko delcev CERN v Švici, ko je leta 1990 razvil protokol izmenjave podatkov, imenovan "hypertekst transfer protocol" oziroma HTTP in poseben način označevanja besedila HTML, ki je omogočil oblikovanje in mrežno povezovanje datotek. V primerjavi z menijskim sistemom programa Gopher je bil Berners-Leejev "brkljalnik" bolj udoben in uporaben. Razvili so prve brkljalnike LineMode, Lynx in druge tekstovne pregledovalnike HTML dokumentov. Toda takrat svetovni splet še ni povsem opravičeval pridevnika "svetovni" (to takrat niti ni bil), namenjen pa je bil predvsem akademski izmenjavi podatkov. Pravo informacijsko revolucijo

pa je leta 1992 povzročil Marc Andressen, z univerze v Illinoisu, ko je pri svojem delu na superračunalniškem centru NCSA s prvim grafičnim pregledovalnikom Mosaic protokol HTTP vizualiziral. Andressen je skupaj z Jimom Clarkom ustanovil Netscape Communications, firmo, ki je na osnovi programske kode za Mosaic razvila danes že kar znameniti brkljalnik Netscape Navigator. Zanimivo je, da se absolutno največja svetovna družba na področju programske opreme Microsoft ni takoj odzvala novim izzivom, ki jih je prinašal razvoj grafičnega brkljalnika in šele z verzijo MS Oken 95, ki vključujejo tudi Microsoft Internet Explorer, delno prevzema primat, ki ga je do takrat skoraj v celoti imel Netscape.

Vemo, kaj se je zgodilo. Na milijonti strežnik na omrežju Internet smo čakali 20 let, na dvomilijontega pa le 1 leto. Večpredstavnost v dokumentih HTML in enostavnost dela z miško sta glavna vzroka za širitev svetovnega spleta. Število novih dokumentov se iz trenutka v trenutek veča in množica informacij se zdi neverjetno kaotična. Ne obstaja noben splošen standard oblikovanja in objavljajanja dokumentov HTML. Določeno vsebino lahko enakovredno predstavljata največji ekspert ali največji laik. Kako zaupati informaciji, ki jo najdemo na Svetovnem spletu? Kako to informacijo tam sploh najti? Kako "sistematizirati" iskanje informacij oziroma spraviti vire informacij v nekakšen red - oblikovati zbirko uporabnih virov informacij? Če želi knjižničar uporabiti svetovni splet kot vir informacij, mora poznati odgovore na vsa ta ključna ključna vprašanja.

3. Kako najti informacijo na svetovnem spletu

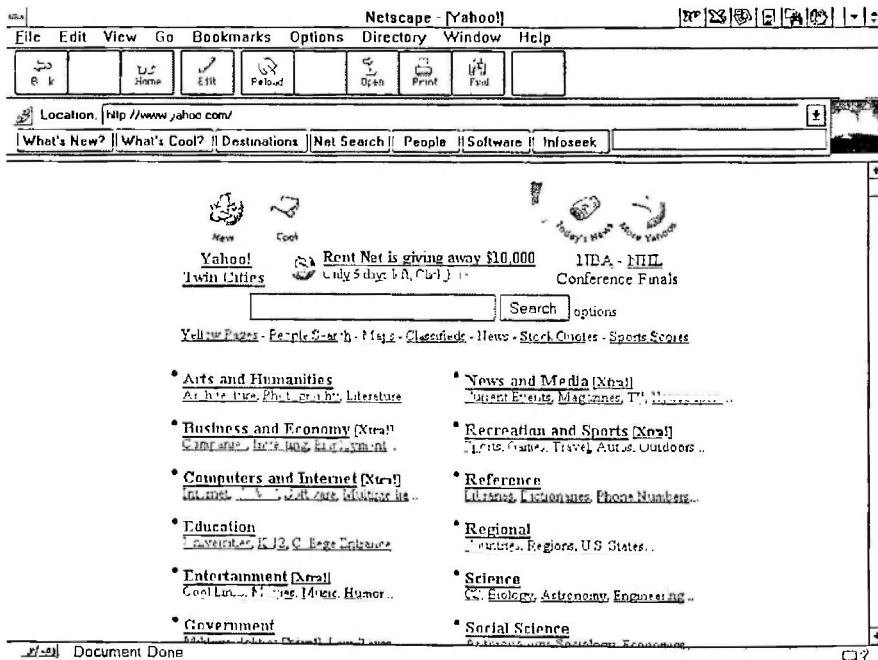
Prva stopnja v oblikovanju zbirke uporabnih virov podatkov je iskanje relevantnih virov na svetovnem spletu. Iskanje po Internetu je lahko glede na prej omenjeni kaotični značaj "zaloge podatkov" frustrirajoče, naporno in neučinkovito, še posebej, če iskanja izvajamo v takoimenovanih "peak hours" oziroma špicah. Po podatkih Arnesa nastopi v Sloveniji prva špica med drugo in tretjo uro popoldne, omrežje pa je najbolj obremenjeno od 20. ure pa do polnoči. Obremenjenost omrežja se razlikuje tudi od kraja do kraja. V Kopru recimo ni popoldanske špice, večerna prej upade, podobno je tudi v Kranju. Na ljubljanskem območju je največ uporabnikov, zato so razmere za delo seveda najtežje, zanimivo pa je, da se podobno dogaja tudi na območju Murske Sobote, kjer je tudi relativno veliko uporabnikov glede na število modemov na vozlišču. Najboljše razmere za delo pa povsod vladajo med 2. in 5. uro ponoči.

V splošnem lahko informacije na svetovnem spletu poiščemo na tri načine:

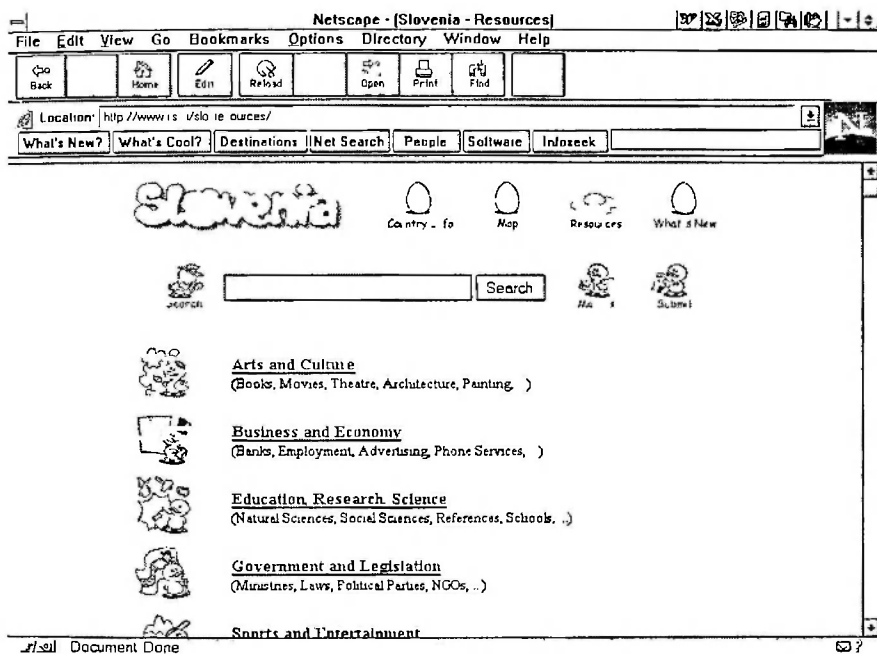
1. z uporabo vsebinskih, kataloško urejenih imenikov,
2. z uporabo iskalnih orodij,
3. kako drugače.

3.1 Uporaba vsebinsko urejenih kataloških imenikov

Uporaba vsebinsko omejenih kataložnih imenikov ima dobre in slabe plati. Lahko so dobro izhodišče, ko izvajamo široke tematske poizvedbe. Organizacija imenikov in podimenikov je hierarhična, kar nam omogoča postopno zoževanje rezultatov. Tako brskanje po katalogih zahteva čas, omogoča pa nam, da osvojimo terminologijo, ki jo uporabljajo avtorji katalogov. Jezik avtorjev oziroma urejevalcev katalogov je ponavadi naravni jezik in se razlikuje od knjižničarske klasifikacijske terminologije za predmetno označevanje. Prav to je zelo šibka točka teh katalogov. Ker jih urejajo različni urejevalci, so si med seboj zelo raznoliki in ne moremo uporabiti enotnega obrazca pri iskanju. Najbolj znan in popoln imenski katalog virov na svetovnem spletu je Yahoo (URL=http://www.yahoo.com).



Podoben kataložni imenik je tudi imenik slovenskih virov na svetovnem spletu (URL <http://www.ijs.si/slo/resources/>), kjer so viri urejeni tudi po organizacijah.



Slika 2: Imenik slovenskih virov na svetovnem spletu

3.2 Uporaba iskalnih orodij

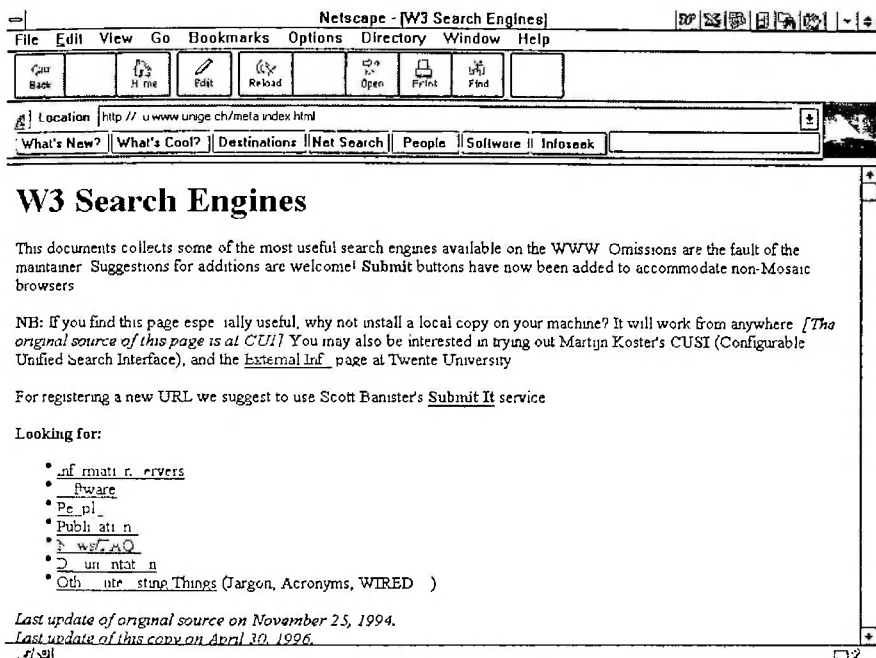
Druga metoda iskanja informacij po svetovnem spletu je vnos iskalnih pojmov ali fraz na določena mesta v HTML dokumentih oziroma straneh, ki imajo funkcijo iskalnega orodja. Iskalna orodja se razlikujejo po načinu iskanja in po širini zaloge podatkov. Nekatera iskalna orodja tako preiskujejo majhne, specifične zbirke na enem strežniku, druga pa različno velike in razpršene zbirke HTML dokumentov po svetovnem spletu. Katero iskalno orodje bomo uporabili, je torej odvisno od:

1. informacije, ki jo iščemo,
2. obsega zbirke, ki jo iskalno orodje pregleduje (ali pregleduje specifične zbirke na enem strežniku ali kar ves svetovni splet, ali pregleduje tudi skupine novic na Usenetu, FTP strežnike...),
3. načina pregledovanja zbirke (nekateri programi preiskujejo samo dele

dokumenta, medtem, ko druga orodja pregledajo ves dokument...),
4. uporabnih možnosti, ki nam jih iskalno orodje nudi (ali ponuja možnost uporabe Boole-ovih operatorjev in omogoča zahtevnejša iskanja, kako razvršča zadetke in jih opisuje...).

Da lahko pravilno zastavimo iskalno strategijo, moramo spoznati osnovna načela, na katerih delujejo iskalna orodja. Obstaja nekaj angleških izrazov za te programe, kot so web robots, spiders, wanderers, worms... Najbolj se je prijel izraz "spiders". Spiders ali pajki so programi, ki križarijo po spletu, vzdržujejo predstavitvene strani in omogočajo iskanje. Programi se med seboj razlikujejo po načinu preiskovanja dokumenta HTML. Nekateri najprej analizirajo naslove dokumentov v zbirkah, preverijo število dostopov in šele nato pregledajo ostale dele dokumenta. Drugi preiskujejo in analizirajo le glavo (header oziroma del dokumenta med paroma znakov ukaza vsebnika HEAD, tretji pregledujejo le določene dele teksta med znakoma ukaza vsebnika Y. Vsako od teh iskalnih orodij mora torej na osnovi iskalne fraze poiskati HTML dokumente po danih kriterijih. Glede na to bi pričakovali, da si bodo rezultati zelo podobni, razlike naj bi nastajale le v hitrosti iskanja. Toda testi revije Internet World so pokazali, da so pri isti iskalni frazi med iskalnimi orodji nastale dramatične vsebinske spremembe, predvsem v rangiranju rezultatov, kar je posledica uporabe različnih kriterijev za določanje stopnje relevantnosti danega zadetka. Veliko zadetkov je bilo manj ali sploh neustreznih.

Najbolj relevantne rezultate je odkril InfoSeek Guide, najbolj obsežne pa Alta Vista.



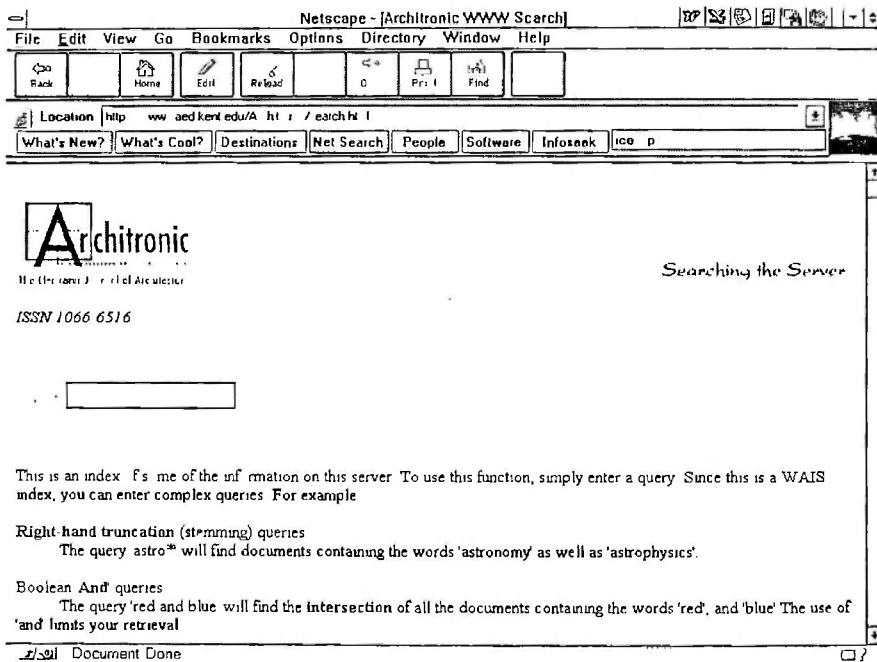
Slika 3: Iskalna orodja so zbrana na strežniku univerze v Ženevi

Poznamo tudi tako imenovana "multiple meta search engines" - iskalna orodja, ki združujejo iskanje več iskalnih orodij hkrati. Imenik teh iskalnih orodij se nahaja na URL <http://ds2.internic.net/tools/meta.html>. Rezultati iskanja s temi iskalci so bolj natančni, predvsem pa manj obsežni.

Velikost svetovnega spleta in vsebinska nestandariziranost HTML dokumentov sta torej glavni oviri pri iskanju informacij na svetovnem spletu. Toda vseeno lahko pričakujemo standard, ki bo v HTML dokumentu imel vlogo nekakšnih ključnih besed. V HTML 3 že obstaja dodatno pojasnilo, v katerem lahko s ključnimi besedami opredelimo HTML dokument, in sicer v okviru ukaza <META>. Ukazu <META> lahko dodamo dodatno pojasnilo NAME="keywords" CONTENT="niz ključnih besed, ločenih z vejico" in tako iskalno orodje posebej opozorimo na ključne besede dokumenta HTML. Ukaz <META> se nahaja v glavi dokumenta, takoj za ukazom vsebnikom <TITLE>. To je pomembno upoštevati pri izdelavi dokumenta HTML, saj bo dokument z dodatnim pojasnilom natančneje opredeljen.

Računalniki, na katerih se nahajajo iskalni programi, imajo seveda izjemne dimenzije. Obstajajo pa tudi relativno enostavni programi, ki nam

omogočajo, da lahko sami na svojem strežniku izdelamo svojo podatkovno zbirko. To je za knjižnice in podobne organizacije zelo pomembno, kajti elektronsko založništvo postaja prav zaradi enostavnosti organizacije svetovnega spleta velika priložnost. Te programe pišemo v jeziku C++ ali v perlu, nahajajo pa se tudi na različnih mestih na svetovnem spletu in jih enostavno prenesemo na svoj računalnik. Primer programa za iskalno orodje, ki ga lahko prenesete na svoj strežnik in ga uporabite kot iskalno orodje za svojo zbirko, je recimo ICE Search Engine, URL <http://www-genome.wi.mit.edu/~čvaidya/ICE/>, avtorja Nadeema N. Vaidye.



Slika 4: Iskalno orodje elektronske revije za arhitekturo Architronic omogoča, da članke iščemo po ključnih besedah.

4. Strategija iskanja po svetovnem spletu

Kakor je svetovni splet vsebinsko pester in raznolik, tako so peestre tudi želje uporabnikov po teh informacijah, zato jih je težko opredeliti v posamezne skupine, vendarle pa lahko rečemo, da smo v CTK svetovni splet kot vir informacij pogosteje kot sicer uporabili v sledečih primerih:

1. iskanje po katalogih tujih knjižnic
2. iskanje po elektronskih knjigarnah in trgovinah z glasbo
3. iskanje statističnih in drugih podatkov različnih vladnih in mednarodnih organizacij (UNICEF, FAO, CIA Factbook, NASA...)
4. iskanje podatkov o določenih organizacijah (naslovi, telefonski imeniki)
5. iskanje poslovnih podatkov
6. iskanje podatkov o univerzah po svetu
7. iskanje podatkov o natečajih (predvsem za študente Šole za arhitekturo)
8. pomoč pri uporabi elektronskih časopisov
9. iskanje in pomoč pri prenosu računalniških programov (shareware, free-ware...)
10. turistične informacije (zemljevidi mest, držav, informacije o hotelih - povpraševanje po tovrstnih informacijah je bolj pogosto pred počitnicami)

Seveda pa je vsebinski spekter informacij, ki smo jih uporabnikom posredovali z uporabo svetovnega spleta še dosti širši.

Pomanjkljiva organizacija znotraj svetovnega spleta nas postavlja pred težavno odločitev izbire ustreznega načina iskanja. Pod ustrezen način iskanja štejemo pravilno izbiro iskalnega orodja in iskani vsebini primerno izbiro iskalne fraze. Oboje hkrati lahko opredelimo kot iskalno strategijo pri iskanju na Svetovnem spletu. Nekaj nasvetov v zvezi s tem:

1. Pred iskanjem poskušajmo osvojiti terminologijo, s katero je iskana vsebina opisana. To lahko naredimo tudi tako, da pregledamo kataloške imenike.
2. Izberemo čim več pojmov, s katerimi bomo iskali.
3. Sprva uporabljamo širše pojme in jih postopoma zožujemo.
4. Podvojimo iskanja na različnih iskalnih orodjih in primerjamo podatke. Ocenimo, katero iskalno orodje je dalo najboljše rezultate. Preglejmo dane rezultate in izločimo manj ustrezne. Preglejmo rezultate na manj uspešnih iskalnih orodjih in tudi tam poskušajmo najti čimbolj ustrezne dokumente. Zapomnimo si, kateri zadetki se podvajajo.
5. Iskanje ponovimo še na katerem od multiplih iskalnih orodij in rezultate primerjajmo.
6. Primerjajmo rezultate, ki smo jih dobili pri pregledovanju kataloških imenikov.

5. Vrednotenje virov informacij na svetovnem spletu

Izdelava dokumenta HTML in postavitev na strežnik je tako enostaven postopek, da lahko upravičeno pričakujemo, da bomo v masi zadetkov, ki jih bomo dobili pri iskanju, naleteli na veliko takih, za katere se po premisleku izkaže, da ne ustrezajo našim željam. Problem svetovnega spleta je

potemtakem tudi to, da ima glede na širino relativno majhno vsebinsko težo, toda tak je njegov značaj in s tem se je pač treba soočiti.

Iskalno orodje poda odgovor na zastavljeno vprašanje v obliki dokumenta HTML, ki je sestavljen iz niza hiperpovezav do dokumentov HTML, ki predstavljajo rezultat iskanja. Kako lahko zaupamo, da so podatki, ki jih najdemo v dobljenih rezultatih, verodostojni, oziroma, da imajo znanstveno ali kakršnokoli že, težo? Da lahko z veliko verjetnostjo govorimo o tem, si moramo o dobljenem dokumentu postaviti nekaj vprašanj:

1. Komu je dokument namenjen? Ali je namenjen akademskim krogom ali pa ima morda le zabavni značaj?
2. Kdo je avtor dokumenta? Posameznik ali organizacija? Če je posameznik, kdo je to? Ali je študent ali izvedenec na določenem področju ali kdo drugi? Če je organizacija, ali je akademska ali komercialna ali druga? Če je komercialna, ali je znana ali ne?
3. Kakšna je frekvenca dostopov?
4. Kdaj je bil dokument nazadnje obnovljen?
5. Kakšen je stil izdelave dokumenta? Ali je tehnično korektno izdelan?
6. Kakšne hiperpovezave vsebuje? Ali vsebuje podobne hiperpovezave, kot ostali, vsebinsko sorodni dokumenti?
7. Ali dokument predstavlja univerzalno informacijo ali je le niz hiperpovezav do drugih, vsebinsko sorodnih virov?
8. Ali je za dostop do dokumenta potrebno geslo?
9. Kakšen je splošni vtis o dokumentu?

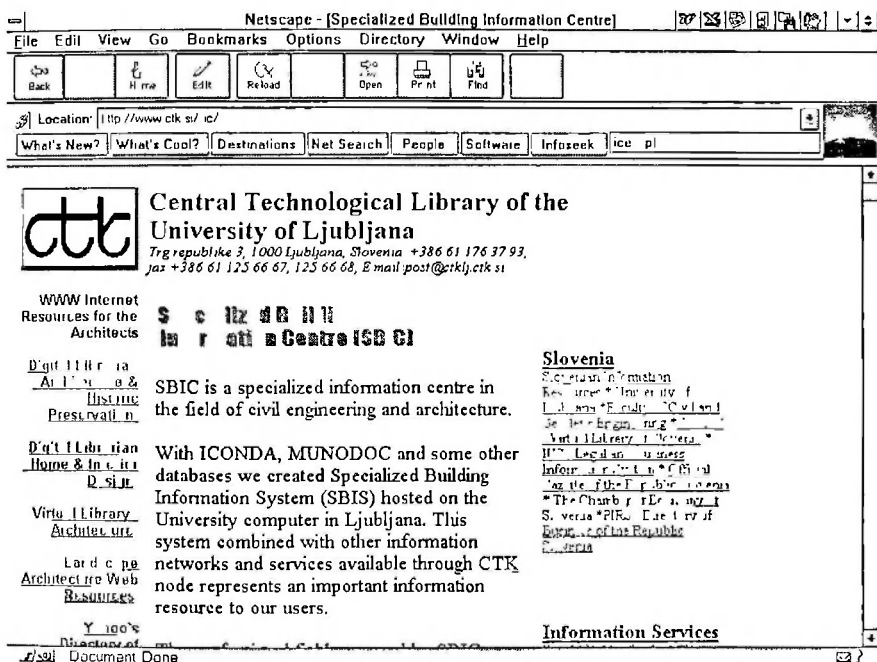
Relevantnost dokumenta, ki ga podajajo iskalna orodja v okviru rezultata, je težko primerjati z resnično ustreznostjo danega zadetka. Podana relevantnost je le matematični seštevek, ki ga naredi program, ko pregleda datoteke in jih primerja z iskalno frazo, medtem ko je vsebinsko ustreznost danega dokumenta vendarle treba preveriti. Toda če se dokument pri večini iskalnih orodij pojavlja na vrhu niza rezultatov v odgovorih, potem lahko sklepamo o resnični ustreznosti našega dokumenta.

Pri tehtanju dobljenih rezultatov je treba imeti torej odprte oči in uporabiti zdrav razum in čimveč predhodnih znanj o iskanih vsebinah.

6. Zaključek - ureditev in predstavitev uporabnih virov informacij na svetovnem spletu

Iskanje po svetovnem spletu si lahko olajšamo tako, da si izdelamo posebne datoteke, v katere beležimo uporabne vire informacij, ki jih srečamo med iskanjem. Za to je potrebno minimalno znanje jezika HTML. Datoteko lahko

uporabljamo interno, le z diska, lahko pa jo postavimo na strežnik in tako tudi drugim omogočimo, da uporabijo naše iskušnje in vloženi trud. Taki indeksi virov informacij pa so lahko tudi deli predstavitvenih strani na svetovnem spletu. Dobro izbrani viri v takem seznamu lahko povedo marsikaj več o našem delu, kot pa predstavitveni tekst sam.



Slika 5: Predstavitvena domača stran Specialnega informacijskega centra za graditeljstvo CTK. Na desni strani je seznam povezav do splošnih informacijskih virov na svetovnem spletu, na levi pa do informacijskih virov iz arhitekture, urbanizma in umetnosti

Iskanje informacij po svetovnem spletu je torej vse prej kot enostavno delo, ki pa sčasoma vendarle preide v rutino. Kljub hitremu porastu števila strežnikov, datotek HTML, predstavitvenih strani, informacij, uporabnikov...na svetovnem spletu, pa se izredno hitro razvijajo tudi pripomočki za iskanje, ki postajajo vedno zmogljivejši in pa predvsem - pametnejši.

7. Literatura

1. **Zorn, Peggy ...** |at al|. "Web searching - tricks of the trade". *Online* 20(1996)3:15-28
2. **Curtois, P. Martin.** "Cool tools for Web searching. An update". *Online* 20(1996)3:29-36
3. **Kimmel, Stacey.** "Robot generated databases on the World Wide Web". *Database* 19(1996)1:41-49
4. **Vendito, Gus.** "Search engine showdown". *Internet World* 7(1996)5:79-96
5. **Nottes R. Greg.** "Forecasts, and Internet Wish List". *Online* 21(1997)1:56-58
6. **Bates, M. Ellen.** "The Internet, part of a professional searcher's toolkit". *Online* 21(1997)1:47-50
7. **Piontek, Sherry; Garlock, Kristen.** "Creating a World Wide Web resource collection." *Internet Research* 6(1996)4: 20-26