

Efektivita informačních služeb ve zdravotnictví a „neviditelný web“ (Health Information Services Efficiency and the „Invisible Web“)

Jiří Menoušek
Nemocnice Louny, p.o., středisko vědeckých informací
svi@nemlouny.cz

INFORUM 2004: 10. konference o profesionálních informačních zdrojích
Praha, 25. - 27.5. 2004

Abstrakt:

Umění hledat v neviditelném webu má vliv na efektivitu informačních služeb ve zdravotnictví. Pojmu „neviditelný web“ použila v roce 1994 poprvé americká odbornice Jill Ellsworthová. Označila tak informace, které nebylo možné na internetu najít pomocí tehdejších vyhledávačů.

Internet je dodnes rozdělen na dvě části – na jedné straně viditelný a „brouzdatelný“ web a na druhé straně až 500krát větší neviditelný, neprůhledný či hluboký web. Právě v něm se nachází i mnoho validních medicínských informací. Kromě technických příčin zde existují další důvody špatné dostupnosti některých informací. Je to stále větší podíl tzv. informačního šumu a také nedostatečná internetová optimalizace dokumentů s odbornými medicínskými informacemi.

Autor vyjmenovává pět bran do neviditelného webu medicíny:

- 1. brána – využití některých zvláštních funkcí klasických vyhledávačů*
- 2. brána – využití specializovaných (medicínských) databází*
- 3. brána – využití specializovaných (medicínských) vyhledávačů*
- 4. brána – využití indexů odkazů na „neviditelné“ dokumenty*
- 5. brána – využití monitorování různých forem on-line diskursu medicínských odborníků (usenetové skupiny, elektronické konference, chaty, atd.)*

Součástí prezentace jsou odkazy na některé zdroje „neviditelných“ medicínských informací.

Abstract:

"Know how" to search on the invisible web, it has the influence on health information services efficiency. The term „invisible web“ was used for the first time in 1994 by an American specialist Jill Ellsworth.. She used this expression for that sort of information which was not accessible by using search engines of that time.

Nowadays the Internet is divided into two parts – on the one hand the visible and surface web and on the other hand as many as 500 times bigger and invisible, opaque or deep web. Exactly on it there are placed a lot of valid medical data .

In addition to technical causes there are many other reasons which make some information difficult to obtain. It has been an increasing rate of noise ratio and insufficient search engine optimization (SEO) of documents with special medical data.

The author gives the list of five gates into the invisible web:

the 1st gate – usage of some functions of common search engines

the 2nd gate – usage of some specialized (medical) databanks

the 3rd gate – usage of some specialized (medical) search engines

the 4th gate – usage of some links indexes for the invisible documents
the 5th gate – usage of the monitorship over several forms of the on-line discourse among medical experts (groups of www users, e-mail conferences, chats, etc.

The part of this presentation are some links to several sources of „invisible“ medical data.

Vliv neviditelného webu na efektivitu informačních služeb ve zdravotnictví

Financování zdravotnictví České republiky prochází v posledních letech hlubokou krizí. Ta se odvíjí od základního faktu, že pacientům je poskytována lékařská péče, srovnatelná v mnoha směrech se standardy vyspělých zemí západní Evropy, systém je ale financován ze zdravotního pojištění, které se odvíjí z „nestandardních“ českých mezd.

Ke kladům českého zdravotnictví patří zatím ještě stále dobrá úroveň know how. Tu zajišťuje slušná úroveň našeho zdravotnického školství, jakož i zatím fungující síť veřejných informačních služeb v medicínských oborech. I v oblasti lékařských knihoven ale dochází ke snižování příspěvků na nákup odborné literatury a informační střediska se proto musí strategicky zaměřit na netradiční zdroje, k nimž patří např. také internet.

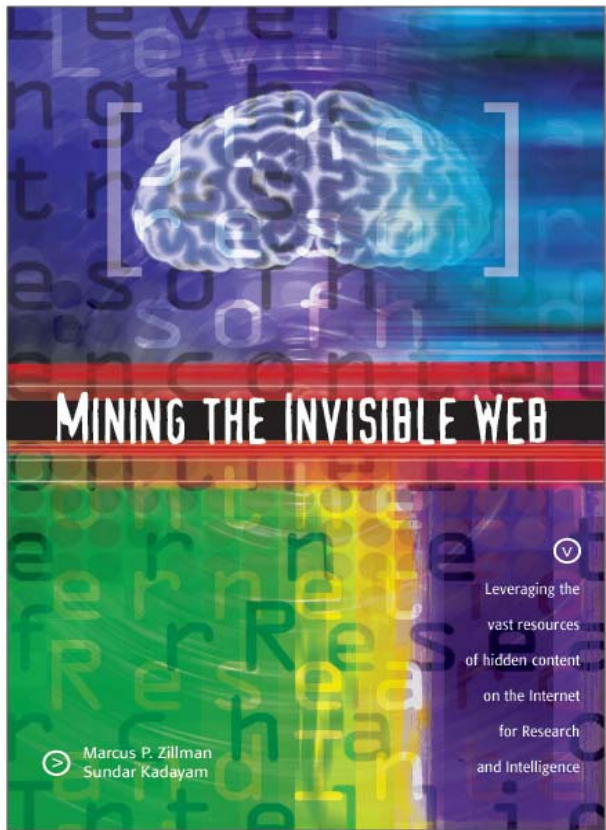
Efektivní využívání internetových zdrojů **není limitováno jen množstvím terminálů a rychlostí připojení**. Důležité je zejména to, aby jak informační pracovníci, tak i zdravotničtí odborníci **uměli na síti příslušné informace rychle najít** a také je kvalitně vyhodnotit.

Jednou z oblastí, která není dostatečně prozkoumána a kde je možné **revolučním způsobem zlepšit efektivitu informačních služeb ve zdravotnictví**, je využívání zdrojů, které nám nabízí oblast tzv. **neviditelného webu**.

Historický úvod

V devadesátých letech minulého století se začala v internetovém prostředí uplatňovat dnešní nejrozšířenější služba sítě - **WWW** (World Wide Web). Projekt byl zahájen v roce 1989 v laboratořích CERN ve Švýcarsku. Ještě roce 1993 tuto službu využívalo na celém světě pouhých 50 serverů. Již o rok později však americká odbornice Jill Ellsworthová (dnes se věnuje **internetovému obchodu**) použila poprvé pojmu „invisible web“ (neviditelný web). Označila tak informace, které nešlo na síti najít pomocí tehdy se rodících klasických vyhledávačů (search engines).

Webový prostor tak byl už z počátku rozdělen do dvou částí. Menší z nich byla pro uživatele běžně dostupná, tedy viditelná a dalo se v ní dobře surfovat (visible web, surface web). Větší část webu však zůstávala stále nepřístupná, temná, neprůhledná - jakoby hluboká a neviditelná (opaque web, deep web, hidden web, invisible web).



Proč je část webu neviditelná? Technologické a jiné příčiny.

Odpověď na tuto otázku je celkem jednoduchá. Většinou proto, že ji „nevidí“ většina klasických vyhledávačů. Jaké jsou příčiny této „slepoty“ i těch nejmodernějších a nejdokonalejších search engines přehledně vypočítal ve svém předloňském **vystoupení** na konferenci **INFORUM** náš přední odborník na tuto problematiku Martin Lhoták. Proto nyní zmíníme jenom tři nejčastější důvody:

- *Ne všechny vyhledávače dokáží indexovat jiné formáty než HTML (Hypertext Markup Language). Jde např. o formáty PDF, RTF, PPT, DOC, XLS, FLASH, streaming media atd.*
- *Zdroj je přístupný pouze po registraci a spider/crawler tuto operaci neumí.*
- *Překážkou jsou dynamicky generované stránky (cgi, javascript, asp a většina stránek s"?" v URL). Stačí, aby spider/crawler zjistil "?" v URL a zastaví se.*

Jednou z dalších příčin je např. časové zpoždění při rešerši. Někteří lidé se totiž ještě stále mylně domnívají, že vyhledávací servery pátrají na webu on-line v reálném čase. Pravda je ale jiná. Search engines pomocí speciálních programů, nazývaných spiders (pavouci) a crawlers (plazíci, lezouni) neustále **automaticky** prohledávají webový prostor a nalezené stránky průběžně ukládají do vlastní gigantické databáze. Teprve z ní, tedy s určitým časovým odstupem, stroj vybírá a pomocí tzv. fuzzy logiky uspořádává odkazy na stránky, které nás aktuálně zajímají. Proto se často setkáváme s tím, že některé nabízené odkazy nelze otevřít. Stránka je buď dočasně nepřístupná, nebo byla (bez tzv. přesměrování) umístěna jinde,

v mnoha případech byl už zdroj dokonce z webové sítě úplně odstraněn. To je však ten menší problém (jak si ještě ukážeme, některé „mrtvé stránky“ můžeme ještě z webu „vydolovat“).

K tomuto výčtu ryze technologických důvodů „neviditelnosti“ některých zdrojů musíme přičíst ještě další faktory. Patří k nim zejména přirozená **expanze množství informací ve webovém prostoru** (která nemá lineární, ale spíše **exponenciální** povahu – počet nových webových stránek roste rychleji, než **rychlost**, jakou jsou vyhledávací stroje schopny stránky zpracovávat).

Dále je to i skutečnost, že se v tomto kynutém informačním knedlíku stále hrozivěji **rozšiřuje podíl tzv. informačního šumu**. Do tohoto balastu, který nepřináší novou kvalitu, patří např. různé multiplicitní reklamy, nevyžádané spamy, tisíckrát omleté a různými kompilátory přežvýkané banality a jiné „plevelné“ informace.

Jinou, tentokrát částečně subjektivní příčinou naší informační slepoty, je používání tzv. internetové optimalizace (**Search Engine Optimization - SEO**). Existují firmy, které dokonale vědí, jak technicky fungují vyhledávače a za různě vysoký poplatek nám zprostředkují dvě služby. V první řadě **upraví** námi dodanou prezentaci tak, aby jinak užitečná, ale v podstatě „nemyslí“ fuzzy logika fulltextových vyhledávačů upřednostnila náš skvost před jinými zdroji. Takto upravené stránky navíc „přihlásí“ do ohromného množství různých rozcestníků a vyhledávačů adresářového typu (mohou jich být řádově až **tisíce**).

The screenshot shows a Microsoft Internet Explorer browser window displaying the website **SubmitStorm.biz**. The browser's address bar contains the URL `http://submitstorm.biz/`. The website's main heading is **SUBMITSTORM.BIZ** in large, bold, red letters. Below the heading, there is a navigation sidebar with four red buttons: **Home**, **Tools**, **Submit**, and **Contact**. The main content area features a descriptive paragraph: "The first and most important step to getting noticed on the Internet is to get listed on search engines. Our service is designed to maximize your exposure on the Internet and increase the number of visitors to your website. We submit your website to over 1200 search engines and directories." Below this text is a dropdown menu labeled "Engine and Directory List-". Further down, there are two "Includes!" sections with logos for **YAHOO!** and **dmoz**. The central promotional message reads: "Submit For Only \$59.99 Plus Free resubmission every month for a year! order now!". To the right of this text is a photograph of a group of people gathered around a laptop. The browser's taskbar at the bottom shows the Start button, several open application windows, and the system tray with the time 13:00.

Hodně nebezpečné (zejména v medicíně, která je na evaluaci zdrojů zvláště citlivá) mohou být např. stránky z dílen většinou komerčně úspěšných organizací, působících v oblasti tzv. alternativní medicíny. Právě takové subjekty využívají služeb „optimalizačních“ internetových firem, které dodají jejich prezentacím „**vědecké balení**“ a speciální webmasterskou úpravu, **napomáhající jejich lepšímu zviditelnění**.

Tak se také stane, že stránky, obsahující ve skutečnosti balast, zavádějící a někdy dokonce smrtelně nebezpečné informace, dostanou přednost před obsahově hodnotnými, ale webmasterem „nešikovně“ zpracovanými a na síti špatně „přihlášenými“ zdroji.

Traduje se, že pokud nám search engine nabídne rozsáhlejší soubor odpovědí, **většina lidí už pátou stránku odkazů ani nečte**, takže i validní a „viditelné“, ale do hlubokého pozadí vytěsňené informace, se pro mnohé naše bližní stávají ve skutečnosti „neviditelnými“ (zde se ovšem nejedná o klasický „invisible web“).

Jak rozsáhlý je neviditelný web a co obsahuje?

Je třeba rovnou říci, že se musíme smířit se dvěma okolnostmi. Invisible web (klasický, neboli „neviditelný“ pro search engines), je za prvé nesrovnatelně větší než jeho „brouzdatelný“ protějšek. Za druhé jako celek obsahuje spíše ty důležitější a hodnotnější informace.

Neoptimističtější odhady hovoří o tom, že viditelný web tvoří maximálně pětinu webového prostoru. Existují ale i studie renomovaných odborníků (např. z organizace BrightPlanet), kteří tvrdí, že **deep web je až 500 krát (!) větší** než jeho viditelná část. Celá internetová síť by se tak vlastně mohla připodobnit k často používanému příkladu plovoucího ledovce, kde nad hladinu ční jen nepatrná viditelná část.

Pokud bychom chtěli zkoumat o co vlastně v oblasti neviditelného webu nejvíce přicházíme, budou to např. validní informace, které jsou uloženy v organizovaných databázích a generují se dynamicky až na základě interakce uživatele se systémem. K nejcennějším uzavřeným zdrojům je zpravidla přístup **chráněn heslem** a databáze jsou plně dostupné jen autorizovaným uživatelům, často pouze na **komerčním základě**. Je nutno zdůraznit, že nákup takových databází je nesmírně drahý. Často se k umožnění přístupu musí složit stovky různých subjektů na základě hromadné uživatelské **multilicence**.

Získání „neviditelné“ informace může znamenat pro některou organizaci velmi významnou konkurenční výhodu. Jinde naopak nedostupnost některých klíčových informací (v medicíně k nim mohou patřit např. lékové interakce nebo toxikologické informace) může rozhodnout o zdraví a v krajním případě třeba i o životě a smrti člověka.

Proč hledat v neviditelném webu?

V zásadě je třeba vycházet z faktu, že i když podle některých (byť částečně intuitivních a nepřesných) studií plných **97,4 procent stránek z oblasti „hlubokého webu“ je veřejně přístupných** (tedy bez jakéhokoli omezení), stejně se nám tam většinou nepodaří najít cestu. Pokus přesto může stát za to. Právě v tomto neviditelném dílu ohromného webového koláče totiž často **najdeme zdroj, který bychom jinak museli za velké peníze získat v té nejmenší, ale zpoplatněné části „deep webu“** (asi 1,6 procent zdrojů má smíšenou povahu - obsahuje volný i placený přístup - a jen **pouhé jediné procento „hlubokého webu“ je limitováno povinným předplatným**).

Traduje se, že co odhodí Skot, to zvedne Holanďan. Pokud jde o bezplatnou exploataci (musíme samozřejmě obětovat vlastní čas a náklady za připojení) internetu, můžeme se celkem oprávněně domnívat, že šetrní Češi předčí oba národy. Nejde ale pouze o

ekonomickou stránku věci. Někdy v oblasti „invisible webu“ můžeme najít informace, za které bychom byli i ochotni dobře zaplatit, klasickými cestami za pomoci běžných vyhledávačů je ovšem stejně nezískáme. Nezbyvá, než se je naučit hledat jinak

Brány do neviditelného webu

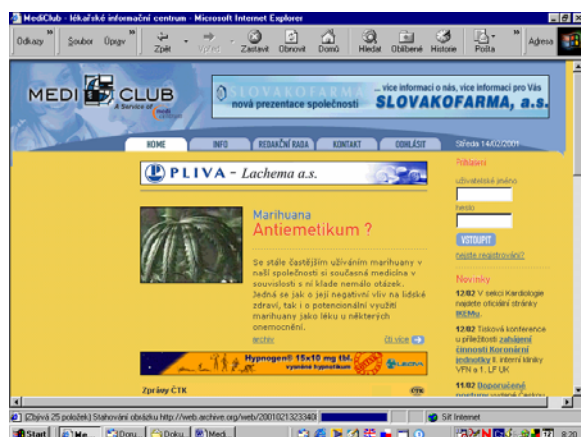
Brána I. - Zvláštní funkce klasických vyhledávačů

Jak si ukážeme, do „neviditelné země“ se můžeme dostat hned několika branami. Nejjednodušší cesta spočívá v důkladném průzkumu možností, které nabízejí klasické vyhledávače. Když zůstaneme u pověstného **Google**, už tady najdeme nejméně dvoje dveře, které nemají mnohé příbuzné vyhledávače. První vchod umožňuje prohlédnout si odkazovanou stránku tak, jak vypadala v době, kdy byla v minulosti vyhledávačem zachycena. Můžeme tak rekonstruovat podobu již „mrtvého“ zdroje, který byl kdysi smazán a dnes ho už na původní adrese nenajdeme. Stačí, když u odkazu, který nám Google nabízí, aktivujeme nápis „**cached**“ (na „českých“ stránkách odkaz [archiv](#)).

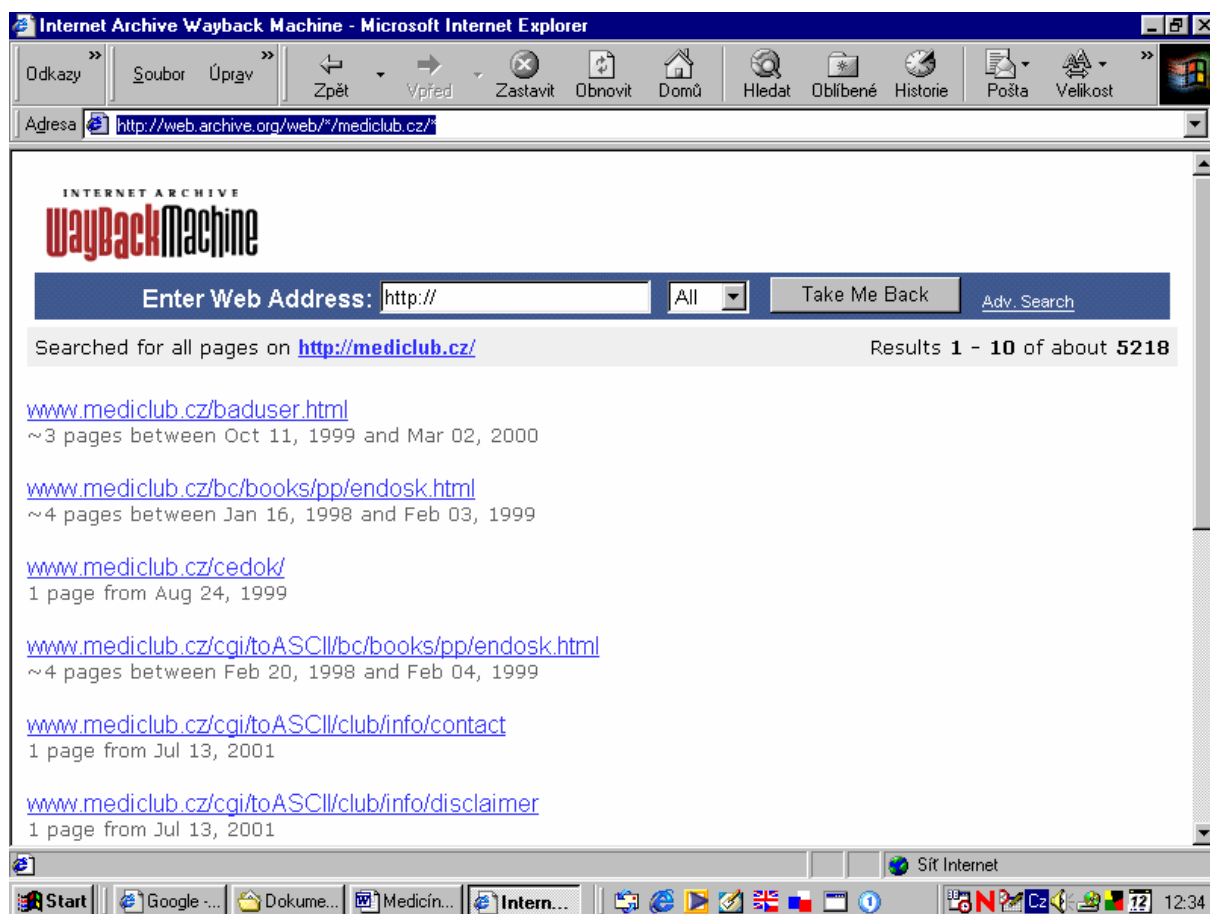
Vyhledávač **Gigablast** (respektive webarchiv **WayBackMachine**) nabízí dokonce celou **databázi** starších verzí stránek.

Můžeme tak např. zjistit, jak vypadaly již dávno smazané a zdánlivě nenávratně ztracené homepage domácího medicínského serveru **MediClub** dne 26. července 1998, 14. února 2001, nebo např. 12. prosince 2002.





Neúnavní šťouralové se dokonce mohou v oblasti hlubokého webu provrtat až k celkem **5218 archivovaným stránkám**, pocházejícím z téhož medicínského portálu, které byly spiders zachyceny v různých časových obdobích.



Další zajímavou cestou je v režimu **rozšířeného vyhledávání** předvolit možnost vyhledávání **některých formátů**, které řada jiných vyhledávačů neumí – např. PDF, powerpointové prezentace a jiné.

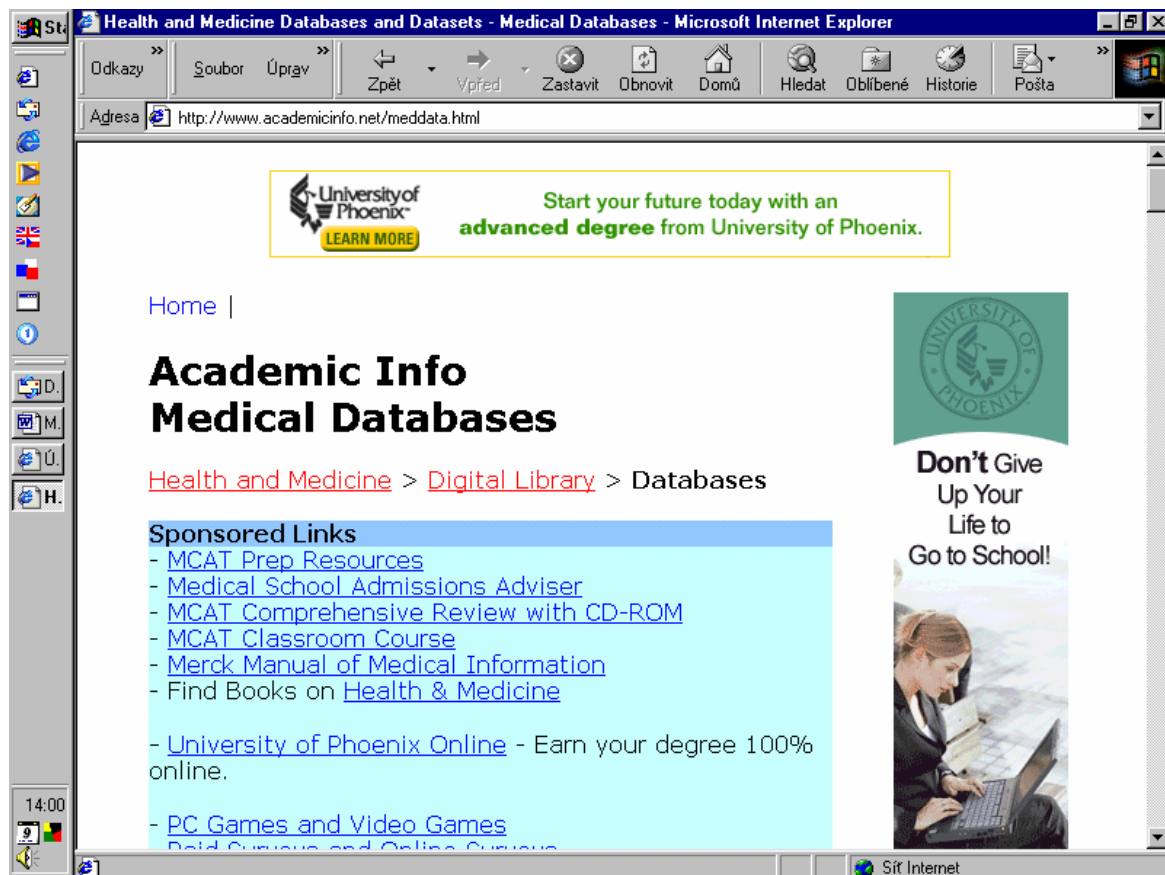
Brána II. - Specializované databáze

Velké množství na první pohled neviditelných informací získáme jednoduše tak, že využijeme některé ze specializovaných databází. Patří sem např. tzv. zpravodajské servery a jejich archivy, různé on-line jízdni řady, telefonní seznamy, specializované bibliografie (např. **elektronických dizertačních prací**), on-line knižní katalogy, obchodní a živnostenské rejstříky, katastr nemovitostí atd. S výběrem vhodného zdroje nám mohou pomoci různě uspořádané **seznamy**, případně univerzální vyhledávače databází, jakými jsou např.:

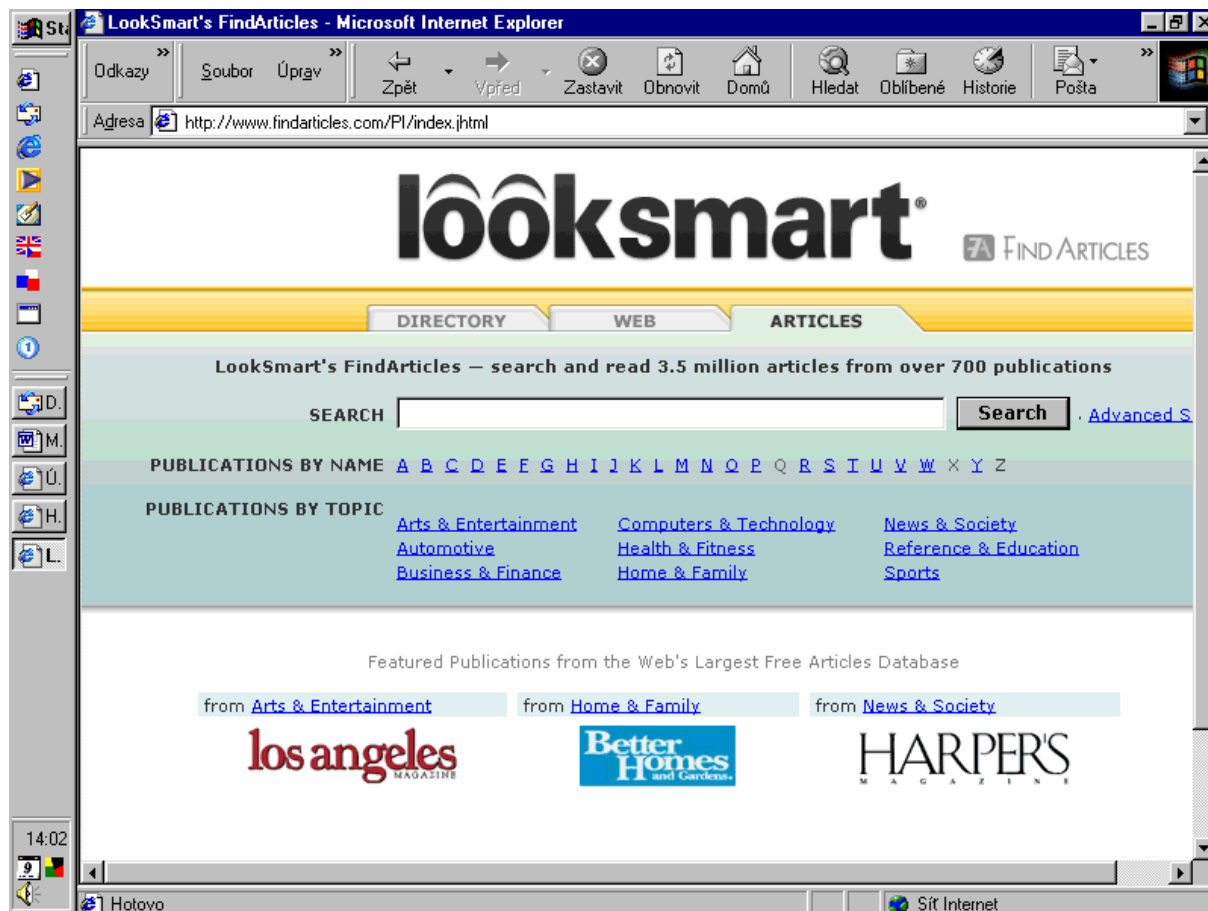
[AcademicInfo](#)

[Infomine](#)

[Librarians Index](#)

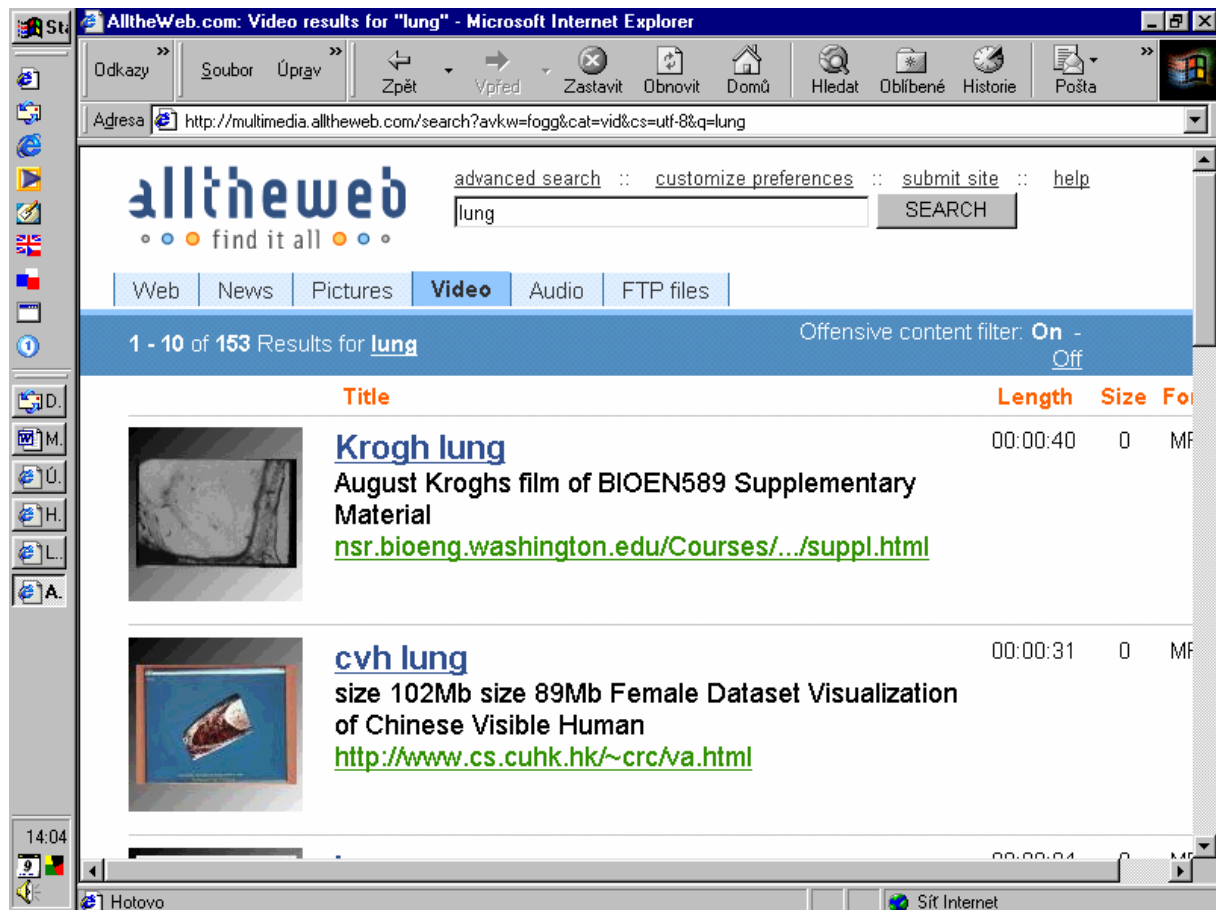


Velmi oblíbené jsou tzv. fulltextové databáze (zde získáme zejména plné texty odborných článků). Většina nejhodnotnějších zdrojů nabízí jen placený přístup, ale najdou se i rozsáhlé free fulltextové databáze typu **FindArticles** a **jiné** zdroje



Pokud zůstaneme u zdravotnictví, k nejužívanějším specializovaným databázím u nás patří např. medicínské bibliografie: **MEDLINE** a **Bibliographia Medica Českoslova**

Další „neviditelné“ informace mohou poskytovat i oblíbené medicínské encyklopedie (viz např. **Healthscout encyclopedia**). Vzhledem k použitým formátům lze mezi obtížněji vyhledatelné zdroje počítat i multimediální dokumenty. Zde pomohou **univerzální vyhledávače**, nebo úzce specializované indexy, např. **videopořadů s medicínskou tematikou**.



Celou plejádou pramenů v oblasti „hlubokého webu“ jsou různé medicínské faktografické databáze. K nejvýznamnějším zdrojům patří organizované databáze léčivých přípravků. Některé, jako např. **RxList** poskytují celé spektrum poznatků, najdou se ale i takové, které se úzce specializují jenom na některý farmakologický problém, např. na lékové interakce. U databáze **Druginteractions** je „neviditelnost“ způsobena také díky dynamicky generovaným stránkám, vznikajícím až na základě interakce vyhledávače s uživatelem, který si sám nastavuje reakce mezi různými vybranými léky (v našem případě mezi Seroquelem a Ataraxem jako jedním z kolekce 846 nabízených reagentů).

Vyjmenovat všechny medicínské specializované databáze není v rozsahu tohoto článku vůbec možné. Je jich na internetu ohromné množství - společným jmenovatelem ale bývá ztížená dostupnost ukrytých informací (které klasické vyhledávače často nevidí a tedy nenajdou).

Nezbývá proto, než si většinou utvořit vlastní terciární informační pramen (jakousi metadatabázi, neboli databázi databází, které se zabývají naší oblíbenou doménou). Potom nám ale stejně nezůstává, než hledat a hledat.

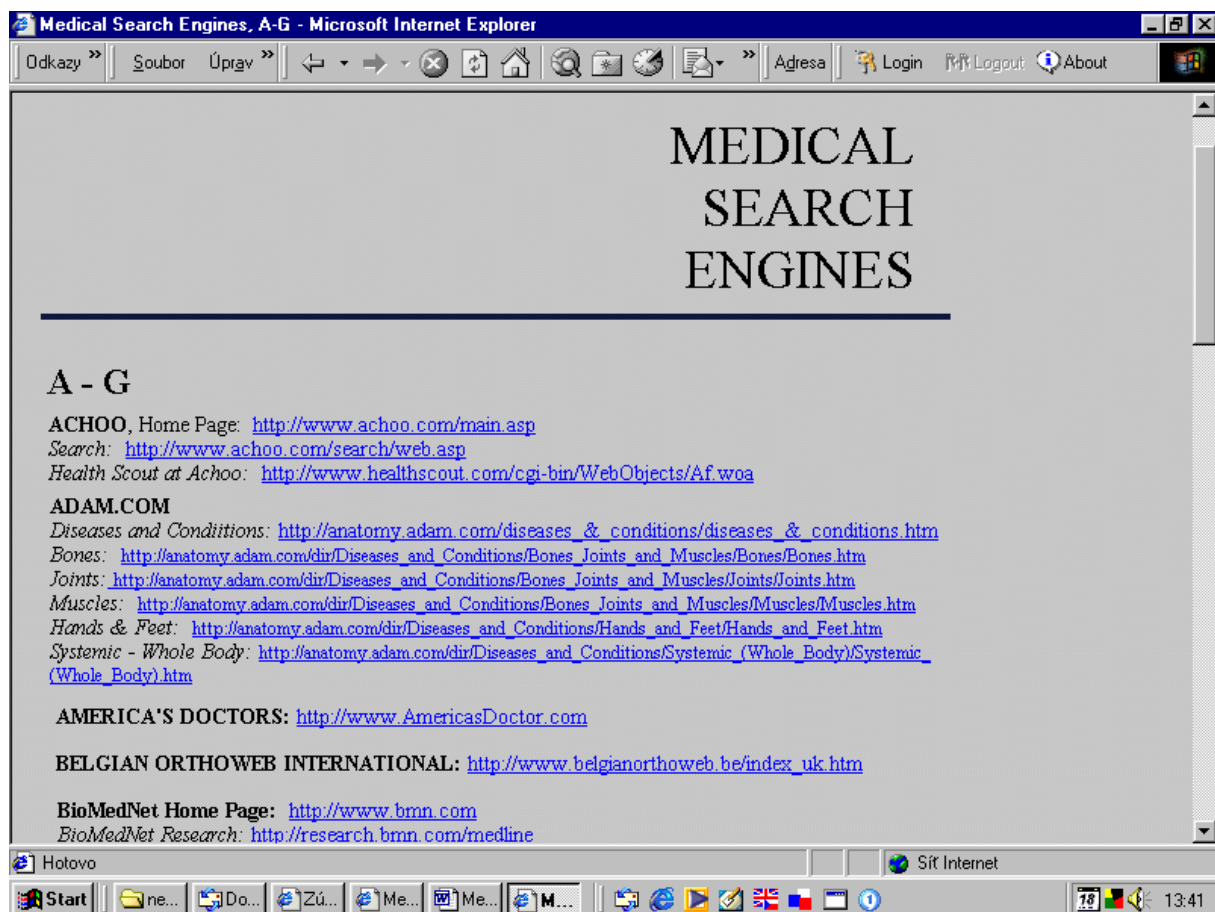
Brána III. - Specializované vyhledávače

Jinou možností, je použít tzv. **specializovaných vyhledávačů**. Jde o search engines, které jsou zaměřeny buď na vyhledávání specializovaných dokumentů (např. u serveru **Scirus** to jsou vědecké články), na materiály z určité geografické oblasti (**Colossus**), nebo přímo na zájmovou doménu, kterou může být becně věda (**SciSeek**), či v našem případě **zdraví** a medicína.

Budeme li např. hledat seznam jihoafrických nemocnic, vyhledávač Colossus nám nabídne přehled 15 regionálních vyhledávačů. Potom stačí najít příslušnou rubriku v jednom z největších jihoafrických search engine ANANZI.



Odkazy na specializované vyhledávače získáme z terciárních informačních pramenů. Jsou to servery, zabývající se buď problematikou vyhledávačů obecně (např. **SearchEngineWatch**), nebo služby, sledující hlavně oborové vyhledávače (mezi jinými i **medicínské**).



Kromě jiného můžeme v těchto terciárních (případně kvartálních) zdrojích najít nejen charakteristiky hledaných služeb, ale i jejich hodnocení (**rating**).

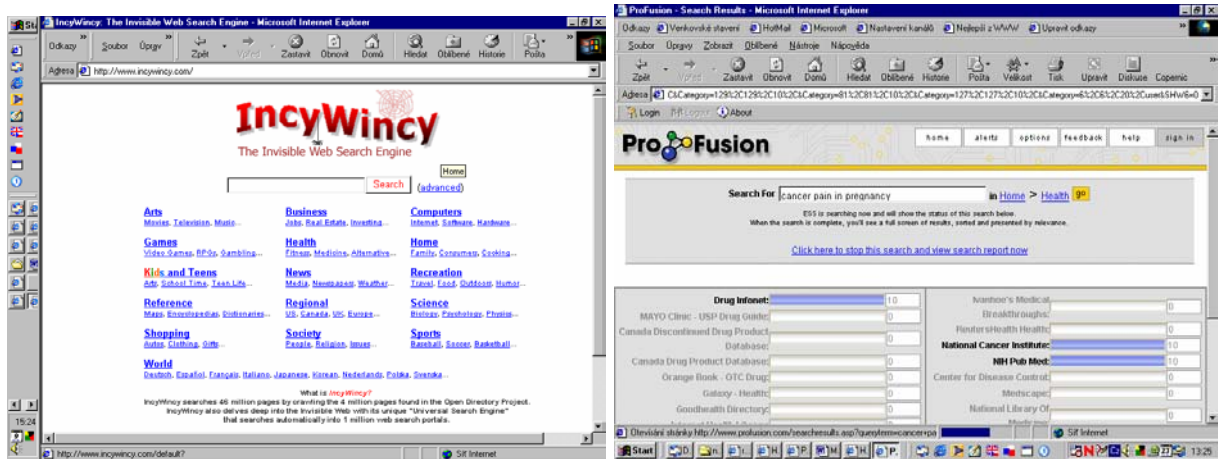
Zde uvádíme několik ukázek indexů univerzálních i specializovaných (např. medicínských) vyhledávačů:

<u>SearchEnginesWorldWide</u>		<u>Awesome Library</u>
<u>Searchability</u>		<u>BeauCoup</u>
<u>SearchCom</u>		
<u>Leiden University</u>		<u>Leiden University(Health)</u>
<u>Fossick</u>		<u>Fossick(Health)</u>
<u>SearchEngineGuide</u>		<u>SearchEngineGuide(Health)</u>
<u>Allsearchengines</u>		<u>Allsearchengines(Health)</u>
<u>SearchBug</u>		<u>SearchBug(Health)</u>
<u>ApplesForHealth</u>		<u>HealthFinder</u>

Brána IV. - Indexy „neviditelných“ stránek

Tak, jako existují klasické soupisy „viditelných“ stránek, najdou se zvláštní multioborové nebo specializované indexy většinou ručně zařazovaných nepřístupných zdrojů. Je to dost pracný způsob akvizice v hloubi sítě uložených pramenů. Tyto mají často nevyrovnanou úroveň a v jedné složce najdeme dosti eklekticky uspořádané zdroje. Výsledek hledání v takovém „bazaru adres“ ale může být docela zajímavý.

<u>Direct Search</u>	<u>Incywincy</u>
<u>CompletePlanet</u>	<u>CompletePlanet/Health/</u>
<u>Invisible-web</u>	<u>Invisible-web/Health/</u>
<u>Profusion</u>	<u>Profusion/Health/</u>



Brána V. - Elektronické skupiny (konference a nástěnky) a příbuzné zdroje

Je to často poslední „brána“, kudy vede cesta k hodně „utajeným“ zdrojům. Víme, že na internetu nepřetržitě probíhají různé elektronické konference a další diskuse účastníků, např. v tzv. newsgroups, nebo on-line chatech. Často jde o diskurs velmi erudovaných odborníků mnoha oborů, rekrutujících se doslova ze všech koutů světa (biofyzik Akeš Bourek v jedné ze svých prací popisuje, že není neobvyklé, když na mezinárodních konferencích najdeme ve vřelém objetí kolegy, kteří se sice nikdy neviděli, přesto se však „důvěrně“ znají z některé elektronické konference).

Můžeme se sami rozhodnout, jestli budeme na těchto zdrojích jenom pasivně „parazitovat“, nebo se sami aktivně zapojíme do diskuse. Obvyklou součástí těchto dialogů, či spíše multilogů (uspořádaných často do monotematických tzv. threads, neboli „nití“) je výměna hypertextových odkazů na různé „neviditelné“ internetové zdroje. Snadno tak můžeme vyřešit problém, se kterým si už musel lámat hlavu někdo před námi. Následuje několik zdrojů, v nichž můžeme začít hledat právě ten náš debatní kroužek.

<u>Google groups</u>	<u>Yahoo groups</u>	<u>AOL/Newsgrups/</u>
<u>MER/SearchForumEngine</u>	<u>CyberFiber</u>	<u>ProfusionDiscussions</u>
<u>TileNet</u>	-	<u>BorlandNewsgroup</u>
<u>Topica</u>		<u>Topica/Health</u>
<u>FAQ Archives</u>	<u>FAQ/Medicine</u>	<u>FAQ/HealthScience</u>
<u>Healthboards</u>	<u>SEZNAM/Zdraví</u>	<u>HealthForums</u>

