II seminarium z cyklu INFOBROKER: *Wyszukiwanie i przetwarzanie cyfrowych informacji,* Centrum Promocji Informatyki, Warszawa 17 kwietnia 2007 r.

Lidia Derfert-Wolf
Biblioteka Główna Uniwersytetu Technologiczno-Przyrodniczego w Bydgoszczy
e-mail: lidka@utp.edu.pl

Odkrywanie niewidzialnych zasobów sieci

I. Niewidzialne zasoby Web

Fragment sieci Internet indeksowany przez standardowe wyszukiwarki (np. Google, Netsprint) określa się jako sieć widoczną / płytką / indeksowaną (ang. *surface web, visible web, indexable web*). Są to zasoby „rozpoznawalne" przez roboty wyszukiwarek (najczęściej statyczne HTML), do których odnośniki znajdują się na innych witrynach albo zostały zgłoszone do baz wyszukiwarek przez użytkowników. Pozostałe zasoby sieci - trudno dostępne dla standardowych wyszukiwarek – nazywane są **ukryty Web** / niewidoczny web / głęboki web (ang. *hidden web, deep web, invisible web*). Terminu *invisible Web* użył po raz pierwszy w 1994 r. Jill Ellsworth dla określenia informacji „niewidzialnych" dla konwencjonalnych wyszukiwarek. Najczęściej cytowane definicje ukrytej sieci to:

- ***Invisible Web*** wg C. Shermana i G. Price'a (2001) -– dostępne w sieci i często bardzo wartościowe strony tekstowe, pliki czy inne informacje, których z przyczyn technicznych bądź innych ograniczeń nie indeksują ogólne wyszukiwarki;
- ***Deep Web*** wg M. K. Bergmana (2001) — strony www tworzone dynamicznie jako wyniki specjalistycznych wyszukiwań w bazach danych.

Badacze sieci ukrytej określili **typy zasobów** tego obszaru Internetu, które obecnie coraz częściej zasilają indeksy wyszukiwarek i tym samym stają się siecią widoczną:

- **zawartość publicznie dostępnych baz danych**

Większość zasobów sieci ukrytej stanowią bazy danych (głównie relacyjne), których zawartości nie jest w stanie spenetrować robot konwencjonalnej wyszukiwarki. Znajdzie on jedynie stronę główną serwisu, gdzie następnym krokiem jest zadanie pytania bazie danych, czego wyszukiwarka już nie potrafi. Tak więc ogromne zasoby (miliardy rekordów, pełnych tekstów publikacji) pozostają „w ukryciu", choć są dostępne w sieci i to często jako strony HTML, ale wygenerowane po zadaniu pytania bezpośrednio bazie danych. W tej grupie zasobów ukrytej sieci mieści się również zdecydowana większość katalogów i innych baz danych tworzonych przez biblioteki, np. bibliografii publikacji pracowników[1]. W tej grupie są również niektóre archiwa gazet i czasopism, słowniki, książki telefoniczne, itp. Ale oczywiście nie jest to regułą. Wszystko zależy od technologii, w której tworzona jest baza danych i jak współpracuje z robotem konkretnej wyszukiwarki.

- **strony i pliki nietekstowe, w innych formatach niż HTML**

Raport Bergmana i praca C. Shermana i G. Price'a powstały w czasie, gdy roboty wyszukiwarek nie radziły sobie z indeksowaniem stron i plików zapisanych w formatach innych niż HTML. Do zasobów ukrytych zaliczano więc wszystkie inne materiały, np.

- dokumenty PDF, Postscript, PHP, doc, xls, ppt, rtf
- witryny czy elementy prezentacji stworzone w technologii Flash czy JavaScript
- pliki multimedialne, nagrania dźwiękowe, obrazy, pliki video
- struktury substancji chemicznych
- programy
- pliki skompresowane (np. *zip, *rar)

[1] Zasoby baz publikacji pracowników tworzonych przez biblioteki np. w Expertusie, ALEPH nie są indeksowane przez Google

Obecnie najlepsze wyszukiwarki bez trudu znajdują strony zapisane w PDF, Postscript, doc, xls, ppt, rtf, o ile spełniają one inne kryteria robotów. Problemem nadal pozostają witryny stworzone w całości lub częściowo w technologii Flash[2], które dla większości robotów są nieczytelne, a więc treści na nich zawarte – często wartościowe – pozostają częściowo „w ukryciu". Roboty nie radzą sobie też plikami graficznymi, dźwiękowymi, video. Dostępne są wprawdzie opcje przeszukiwania grafiki, audio czy video, ale np. robot Google nie rozpoznaje tekstu zawartego w grafice – wyszuka obrazek tylko wtedy, gdy nazwa pliku lub tekst w jego pobliżu w kodzie źródłowym odpowiada naszemu zapytaniu.

- **strony, do których nie prowadzą odsyłacze z innych witryn**

Zgodnie z jedną z zasad działania robotów wyszukiwarek – indeksowania stron, do których prowadzi przynajmniej jeden link z innej strony - do indeksów wyszukiwarek nie zostaną włączone strony, do których brak odsyłaczy z innych stron.

- **strony wyłączone z procesu indeksacji przez twórców**

Niektóre serwisy (w tym komercyjne bazy danych) wymagają płatnej rejestracji bądź są dostępne dla subskrybentów po dokonaniu autoryzacji, więc ze względów oczywistych takie zasoby muszą być wyłączone z procesu indeksacji przez wyszukiwarki. Roboty nie są również w stanie, z powodów technicznych, zalogować się do serwisów wymagających bezpłatnej rejestracji. Te dwa rodzaje zasobów tworzą tzw. sieć zastrzeżoną, której pokaźną częścią są cenne naukowe bazy danych, np. Inspec, EBSCO. Druga grupa stron wyłączonych z indeksacji to te, których twórcy zabronili robotom indeksowania treści dokumentów i zamieścili pliki robots.txt w kodzie źródłowym strony.

- **strony i pliki tworzone dynamicznie lub w czasie rzeczywistym**

Strony generowane dynamicznie, w ramach zazwyczaj dużych serwisów internetowych tworzonych przy pomocy np. technologii ASP czy PHP, powstają po zainicjowaniu pewnej „akcji" przez użytkownika, np. zadaniu pytania bazie danych albo wyszukiwarce serwisu, wypełnieniu formularza czy ustawianiu własnych preferencji. Jest to więc informacja generowana „w locie" i tworzona przez specjalne skrypty, zgodnie z potrzebami konkretnego użytkownika strony i niekoniecznie istotna dla innego użytkownika. Adres URL stron tworzonych dynamicznie jest zazwyczaj bardzo długi i zawiera ? lub &. Nie każdy robot wyszukiwarki odnajdzie taką stronę, a te które znajdują często nie zamieszczają ich w swoich indeksach. Kolejnym problemem dla standardowych wyszukiwarek są strony tworzone w czasie rzeczywistym, których zawartość zmienia się prawie każdego dnia, np. notowania giełdowe, prognozy pogody, rozkłady lotów. Roboty często celowo rezygnują z ich indeksowania, gdyż są to dane "ulotne", krótkotrwałe, a przede wszystkim w wielkich ilościach.

Co jeszcze jest tak naprawdę „widoczne", ale „ukrywa się"?

Coraz więcej z w/w zasobów jest i będzie indeksowanych przez standardowe wyszukiwarki. Pozostanie natomiast, a wręcz będzie narastał, problem przeładowania informacji. Rezultaty wyszukiwania np. w Google to zazwyczaj setki, tysiące witryn. Użytkownik zagląda do pierwszych kilkudziesięciu, tymczasem to, co go najbardziej interesuje może być dużo „głębiej". Kolejnym problemem jest umiejętność zadawania pytań. Programy wyszukiwawcze są coraz bardziej „inteligentne", jednak wyszukanie specyficznej informacji ciągle wymaga sformułowania precyzyjnego pytania i korzystania z formularzy zaawansowanych, co nie jest powszechne. Ponadto przyzwyczajenie do jednej wyszukiwarki – „mit" Google – i pomijanie innych narzędzi, które rejestrują zasoby „widzialne", np. katalogów tematycznych czy wyszukiwarek specjalnych (osób, grafik, blogów itp.), powoduje otrzymywanie w rezultatach wyszukiwania sporo zdublowanych i zbędnych informacji.

Jak wielkie są zasoby ukryte i jakie to rodzaje informacji?

Powszechnie wiadomo, że nie można precyzyjnie zliczyć wszystkich stron WWW, ani tych, które „widzą" wyszukiwarki. Mimo to, co pewien czas podejmowane są badania, w których naukowcy starają się oszacować **wielkość sieci płytkiej i głębokiej**. Najczęściej cytowana praca, na której opierało się wiele następnych to wyniki badań M. K. Bergmana dla Bright Planet z 2001 r., wg których:

- *deep web* jest ok. 400-550 razy większy niż zasoby indeksowane przez wyszukiwarki i liczy ok. 550 mld dokumentów (7.500 terabajtów);
- 60 największych baz danych ukrytego Internetu zawiera 750 TB danych;

[2] Np. robot Gooru - jeśli pierwsza strona serwisu jest w całości wykonana we Flash'u i do kolejnych podstron nie prowadzi żaden link w formacie HTML robot nie znajdzie pozostałych stron w serwisie.

- ok. 95% zasobów "ukrytych" jest dostępnych publicznie, bezpłatnie;
- ponad połowa to tematyczne bazy danych;
- zasoby *deep web* są odwiedzane przez użytkowników o 50% częściej niż zasoby „płytkie".

C. Sherman podał w 2001 r. nieco inne wyniki badań i oszacował ukryty web jako ok. 2-50 razy większy niż zasoby indeksowane przez wyszukiwarki. D. Lewandowski i P. Mayr, P uznali w 2006 r. szacunki M. K. Bergmana za bardzo zawyżone z powodu błędu statystycznego (korzystania ze średniej zamiast mediany) oraz liczenia rozmiaru baz danych w GB zamiast rekordach. Sami oszacowali rozmiar „naukowej sieci ukrytej" na ok. 20-100 mld dokumentów. A. Gulli i A. Signorini opublikowali w 2005 r. wyniki badań nad siecią indeksowaną przez wyszukiwarki (*surface Web*), szacując jej wielkość na ok. 11,5 mld stron. Z tego zbioru 9,36 mld stron jest dostępnych w indeksach czterech największych wyszukiwarek (Google, Yahoo, MSN, Ask). Wykazali również, że przeciętnie ok. 70% zasobów *surface web* można było uzyskać za pomocą w/w wyszukiwarek (np. Google 76%), a część wspólna indeksów czterech w/w wyszukiwarek wynosi 2,7 mld stron (28,85%). Standardowe wyszukiwarki uniwersalne ciągle doskonalą mechanizmy swoich robotów i powiększają dzieki temu swoje bazy/indeksy, pomniejszając przy tym rozmiar ukrytej sieci. Jednak na razie żadna z nich, nie wyłączając Gogle, nie jest w stanie zindeksować całego Web.

Wymienione wcześniej typy ukrytych zasobów sieci wyszczególniono w punktu widzenia technicznego, technologicznego bądź prawnego, biorąc pod uwagę rodzaje formatów i metody tworzenia zasobów. Jakie natomiast rodzaje informacji kryją się w sieci pozornie niewidzialnej? Autorzy wielu publikacji na ten temat zgodnie podkreślają, że większość stanowią bardzo wartościowe materiały, a pośród nich:

- publikacje i raporty naukowe, dysertacje (pełne teksty lub abstrakty);
- artykuły z gazet i czasopism (pełne teksty lub abstrakty);
- dokumenty rządowe;
- archiwa materiałów źródłowych i referencyjnych;
- zasoby biblioteczne (katalogi, zbiory digitalizowane, publikacje pracowników uczelni);
- niektóre repozytoria Open Access;
- szara literatura;
- dane, wzory, grafiki;
- słowniki i encyklopedie, bazy teleadresowe;
- i wiele, wiele innych.

W tej grupie dokumentów przeważająca część to materiały przydatne dla nauki i edukacji. D. Lewandowski i P. Mayr (2006) zaproponowali termin „**Naukowa Sieć Niewidzialna**" dla określenia baz danych i kolekcji o istotnym znaczeniu dla środowisk naukowych, bibliotekarzy i specjalistów informacji, a nie wyszukiwalnych przez standardowe wyszukiwarki. Naukowa Sieć Niewidzialna zawiera literaturę (np. artykuły, rozprawy, raporty, książki), dane (np. dane z badań), materiały wyłącznie online (np. dokumenty Open Access). Głównymi dostarczycielami tych zasobów są: twórcy i sprzedawcy baz danych i innych serwisów (np. document delivery), biblioteki, wydawcy komercyjni, uczelnie, instytucje i stowarzyszenia naukowe.

II. Sposoby „odkrywania" niewidzialnych zasobów sieci

Odkrywanie zasobów ukrytych to z jednej strony obejmowanie coraz większych obszarów sieci przez wyszukiwarki, z drugiej zaś umiejętne korzystanie z zasobów pozostających poza zasięgiem wyszukiwarek. Pierwszy problem może być rozwiązywany przez rozwijanie możliwości robotów wyszukiwarek, ale również podejmowanie ich współpracy z dostawcami informacji, dotąd zastrzeżonej dla ogółu użytkowników sieci. Drugi problem to odpowiedź na pytanie „dlaczego warto nie poprzestawać na Google?" w wyszukiwaniu informacji. Wspomniano już, że wszystkich zasobów Web nie da się do końca „uwidocznić". Ponadto żadna wyszukiwarka nie obejmuje swym zasięgiem całej sieci „widocznej". Rezultaty wyszukiwania często są zdublowane – te same informacje pochodzą z różnych serwisów. Odmienną sytuację mamy w serwisach *deep web*, gdzie informacje rzadko się powielają i charakteryzują się wysoką jakością. Warto zatem podejmować pewne działania zmierzające z jednej strony do wzajemnej „współpracy" tworzonych zasobów (szczególnie bazodanowych i repozytoriów OA) oraz ich „współpracy" ze standardowymi wyszukiwarkami, z drugiej – uświadamiania użytkownikom internetu istnienia ukrytego web. Starania te powinni podejmować zarówno twórcy baz danych, wydawcy, biblioteki (razem!), ale też twórcy wyszukiwarek.

1. Specjalne usługi standardowych wyszukiwarek, współpraca z wydawcami czasopism oraz twórcami baz danych i katalogów

Największe wyszukiwarki oferują odrębne usługi, dzięki którym można dotrzeć do część sieci ukrytej. Są to najczęściej odnośniki do publikacji naukowych, rządowych, artykułów z czasopism czy zasobów baz danych. Specjalistyczna wyszukiwarka Google Scholar[3] rejestruje pełne teksty lub abstrakty prac naukowych (artykułów, raportów, książek), w tym linki do publikacji z wydawnictw komercyjnych. Publicznie dostępne są pełne teksty publikacji funkcjonujących w sieci bez ograniczeń, jak również informacje bibliograficzne oraz abstrakty tekstów z baz komercyjnych. Pełne teksty tych ostatnich dostępne są dla subskrybentów na zasadzie współpracy Google z wydawcami i dostawcami baz, np. EBSCO. W Google Scholar jest też możliwość sprawdzenia dostępności książek w wielu bibliotekach (w tym NUKAT – opcja „Find In NUKAT"), dzięki współpracy z WorldCat OCLC czy zamówienia kopii w artykułu w British Library. Z innych usług Google warto wymienić Google Books[4], współpracujący również z Worldcat OCLC (opcja "Find this book in a library") oraz Google Patent Serach[5] (7 mln pełnych tekstów opisów patenowych USA).

Wyszukiwarka Live Search[6] (MSN) uruchomiła usługę *Academic*, która umożliwia wyszukiwanie artykułów i innych publikacji naukowych, m.in. z repozytoriów Open Access. Pełne teksty publikacji mogą uzyskać użytkownicy z instytucji opłacających dostęp do serwisów czasopism elektronicznych lub baz danych, np. EBSCO, Elsevier, Wiley. Wymagane jest jednak uprzednie zgłoszenie instytucji do Live Serach i dokonanie odpowiednich ustawień w wyszukiwarce.

Kolejny potentat na rynku wyszukiwarek – Yahoo! – oferuje usługę *Search Subscriptions*[7], dzięki której subskrybenci różnych baz danych mogą przeszukiwać jednocześnie wszystkie zasoby, do których mają uprawnienia, np. LexisNexis, IEEE publications, Consumer Reports.

Ponadto dostawcy, np. IEEE Xplore, „ujawniają" dla standardowego Google informacje ze swoich baz do poziomu np. opisu bibliograficznego, a reszta (abstrakt, pełen tekst) jest zastrzeżona. Jeśli na taki odnośnik trafi użytkownik, którego instytucja opłaca dostęp do pełnych tekstów, natychmiast otrzymuje ten tekst na ekranie swojego komputera, na podstawie kontroli numerów IP.

2. Stosowanie odpowiednich standardów i innych rozwiązań technicznych przez twórców zasobów sieciowych

To zagadnienie dotyczy przede wszystkim twórców bezpłatnych baz danych i dynamicznych stron internetowych, których prace zmierzają w kierunku uwidaczniania swoich zasobów standardowym wyszukiwarkom. Jednym z rozwiązań jest konwersja bazy danych do regularnych stron HTML, np. w Amazon.com, każdy rekord jest konwertowany do HTML i „widoczny" dla robotów wyszukiwarek. Można to sprawdzić porównując URL konkretnego rekordu wyszukanego bezpośrednio w Amazon.com i potem tego samego tytułu wyszukanego w Google.

Stosowanie protokołów Z39.50 i/lub OAI-PMH[8] umożliwia jednoczesnego przeszukiwanie (*cross search*) wielu serwisów i baz danych, tych ukrytych i dostępnych dla wyszukiwarek. Doskonałym przykładem jest brytyjski serwis TechXtra[9] ukierunkowany na inżynierię i technikę, przeszukujący 4 mln rekordów z 30 baz danych, serwisów tematycznych, repozytoriów OA. Innym ciekawym przykładem jest E-Print Network[10] - serwis przeszukujący repozytoria e-printów z zakresu nauki i techniki. Protokół OAI-PMH pozwala również na jednoczesne wyszukiwanie w zasobach bibliotek cyfrowych oraz ich uwidaczniania w wyszukiwarkach (Heliński, M. i inni 2005). Dzięki temu zasoby bibliotek polskich stosujących oprogramowanie dLibra[11] mogą być przeszukiwane wspólnie (opcja „Przeszukaj zdalne biblioteki" na stronie którejkolwiek z nich), są „widoczne" w Google i wyszukiwarkach specjalistycznych, np. OAIster.

---

[3] http://scholar.google.com
[4] http://books.google.com/
[5] http://www.google.com/patents
[6] http://search.live.com/
[7] http://search.yahoo.com/subscriptions
[8] Open Archives Initiative Protocol for Metadata Harvesting. Zob. więcej na stronie dLibra: http://dlibra.psnc.pl/index.php?option=com_content&task=view&id=62&Itemid=62&lang=pl
[9] http://www.techxtra.ac.uk/
[10] http://eprints.osti.gov/
[11] http://dlibra.psnc.pl/ (zob. Wdrożenia)

3. Tworzenie specjalistycznych wyszukiwarek i multiwyszukiwarek[12]

Wyszukiwarki specjalistyczne umożliwiają wyszukiwanie w ramach określonej dziedziny (np. technika, sztuka), formatu plików (np. dźwiękowych, video, *pdf), rodzaju informacji (np. rządowe) bądź są przeznaczone dla okreslonej grupy odbiorców (np. wyszukiwarki naukowe). Mogą też łączyć kilka z tych kryteriów. Dobrym przykładem jest Internet Archive - biblioteka internetowa dla naukowców, oferująca dostęp do kolekcji z zakresu nauk humanistycznych w postaci cyfrowej, w tym tekstów Open-Access, plików audio, filmów i programów. Doskonałą wyszukiwarką plików graficznych jest Picsearch. Z wyszukiwarek naukowych najpopularniejszymi są Scirus, CiteSeer, Find Articles. Na uwagę zasługują również wyszukiwarki w repozytoriach i czasopismach Open Access, np. OAIster - katalog i wyszukiwarka zasobów bibliotek cyfrowych różnych instytucji, repozytoriów instytucjonalnych i czasopism elektronicznych, Open Access DOAJ - katalog czasopism bezpłatnie dostępnych w sieci czy OpenDOAR - wyszukiwarka akademickich repozytoriów.

Druga grupa specjalistycznych narzędzi wyszukujących to różnego rodzaju multiwyszukiwarki kierujące pytanie użytkownika do wielu wyszukiwarek, baz danych i innych serwisów, w tym zasobów *deep web*. Interesująca jest multiwyszukiwarka GoshMe, przeszukująca indeksy wyszukiwarek ogólnych i specjalistycznych. Z innych na uwagę zasługują Turbo10, INCYWINCY, Trovando. Typowe katalogi zasobów ukrytych, w większości baz danych, to Geniusfind i CompletePlanet.

4. Zintegrowane przeszukiwanie elektronicznych zasobów bibliotek

Przy rosnącej liczbie zasobów elektronicznych dostępnych dla użytkowników jednej instytucji, np. biblioteki, materiały te stają się „niewidoczne" dla użytkowników, bo …. jest ich bardzo dużo i linki do nich zamieszczane są w różnych działach, pod nazwami nie zawsze czytelnymi. Powstają wobec tego programy ułatwiające poruszania się po witrynach bibliotecznych, coraz bardziej „obładowanych" informacjami i dostępami do wielu katalogów, baz danych, pełnych tekstów artykułów, bibliotek cyfrowych i innych repozytoriów. Najlepszym rozwiązaniem jest przeszukiwanie wszystkich zasobów elektronicznych danej biblioteki (lokalnych i zdalnych) przy pomocy jednego interfejsu – jedno pytanie do kilku bądź wszystkich serwisów. Oczywiście dostęp do zasobów komercyjnych mają wtedy wyłącznie zarejestrowani użytkownicy. Istnieje wiele rozwiązań typu *federated searching*[13]. Przykładem jest serwis RUG Combine biblioteki Uniwersytetu w Groningen[14], AquaBrowser Queens Library[15] czy BASE – zintegrowana wyszukiwarka katalogu biblioteki uniwersyteckiej w Bielefeld oraz ok. 160 serwisów OA[16].

5. Tworzenie katalogów tematycznych, w tym serwisów o kontrolowanej jakości *subject gateway*[17]

Tworzenie katalogów tematycznych pozostaje w gestii ludzi, a maszyny jedynie tę działalność wspierają. Katalogi wywodzą się ze zwykłych wykazów przydatnych linków, zazwyczaj wyselekcjonowanych dla określonej grupy odbiorów. Najcenniejsze dla użytkowników są serwisy typu *subject gateways* (Derfert-Wolf, L. 2004) czyli serwisy tematyczne o kontrolowanej jakości[18]. Są to dziedzinowe przewodniki (miejsca startowe) po zasobach internetowych, uporządkowane według kategorii tematycznych. Zasoby są selekcjonowane, oceniane, opisywane i katalogowane przez bibliotekarzy lub ekspertów z danej dziedziny. Linki zgromadzone w tych serwisach dobiera się zgodnie z oficjalnie opublikowaną listą kryteriów oceny jakości i opisuje według powszechnie stosowanych systemów klasyfikacyjnych. Natomiast gwarancją regularnej aktualizacji i kontroli linków jest profesjonalne zarządzanie ich kolekcją. *Subject gateways* dają w efekcie zorganizowany zbiór (bazę danych) linków do źródeł o kontrolowanej jakości, przeszukiwany według słów kluczowych i przeglądany wg kategorii tematycznych. Rezultaty zawierają standardowe opisy - zestaw metadanych.

Cechy charakterystyczne serwisów tematycznych o kontrolowanej jakości to:

1) serwis online kierujący do wielu innych serwisów lub dokumentów w sieci;
2) dobór źródeł jest twórczym procesem zgodnym z opublikowanymi kryteriami jakości (wyklucza się np. dobór na podstawie automatycznie mierzonej popularności źródeł);

[12] Adresy URL i krótkie opisy omawianych w tym dziale serwisów podano w Tab. 1
[13] Lista programów m.in. na stronie http://www.loc.gov/catdir/lcpaig/portalproducts.html.
[14] http://www.rug.nl/bibliotheek/
[15] http://aqua.queenslibrary.org/
[16] http://www.base-search.net/
[17] Adresy URL i krótkie opisy omawianych w tym dziale serwisów podano w Tab. 1
[18] DESIRE Information Gateways Handbook http://www.desire.org/handbook/

3) opis źródła (od krótkiej adnotacji do recenzji) jest również procesem twórczym (wyklucza się np. kopiowanie streszczeń bezpośrednio z witryn polecanych stron/serwisów). Dobrym, ale nie koniecznym kryterium jest dodawanie słów kluczowych lub deskryptorów.
4) głęboka struktura tematyczna bądź system klasyfikacji, powstały również w trakcie procesu intelektualnego i służący do przeglądania serwisu (wyklucza się listę linków bez żadnej struktury).
5) manualnie (przynajmniej częściowo) tworzone metadane (opis bibliograficzny) dla każdego źródła.

Serwisy typu *subject gateways* można dzielić według tematyki, języka, terytorium, współpracy i wielu innych. Na świecie istnieje sporo różnych inicjatyw, realizowanych głównie przez środowiska akademickie, w tym bibliotekarzy. Do największych katalogów rejestrujących źródła z różnych dziedzin zaliczyć należy BUBL, Infomine, Scout Archives. Z serwisów tematycznych tworzonych przez kilka instytucji naukowych na uwagę zasługują Intute i Vascoda. Adresy URL i krótkie opisy tych katalogów podano w Tab. 1.

6. Szkolenie użytkowników sieci

Wszystkie omówione wyżej sposoby odkrywania niewidzialnych zasobów sieciowych nie przyniosą oczekiwanych efektów, jeśli użytkownicy będą mieli niewielką wiedzę na ich temat. Zatem kolejną metodą docierania do tych, niekiedy tylko pozornie, ukrytych zasobów jest uświadamianie istnienia *deep Web* i wskazywanie sposobów poszukiwań tych źródeł w sieci. Najlepszą metodą upowszechniania wiedzy o ukrytych zasobach Web jest prezentowanie przykładowych stron WWW pozostających poza zasięgiem standardowych wyszukiwarek, np. Google. Reakcją jest wtedy pytanie „jeśli nie w Google to gdzie?". Odpowiedzią powinien być przegląd najważniejszych baz danych, multiwyszukiwarek, narzędzi specjalistycznych, katalogów tematycznych przydatnych dla szkolonej grupy. Należy podkreślać, że nie warto poprzestawać na Google w wyszukiwaniu wartościowych informacji, a zasobów standardowo „widzialnych" przez Google lepiej nieraz szukać gdzie indziej - wyszukiwanie będzie bardziej efektywne i da lepsze rezultaty. Standardowe wyszukiwarki przydadzą się natomiast w dotarciu do baz danych i innych serwisów *deep web*. Warunkiem jest precyzyjne sformułowanie pytania, która to umiejętność wraz z korzystaniem z formularzy wyszukiwania zaawansowanego przyda się w ogóle w przeszukiwaniu zasobów sieci i odkrywaniu cennych materiałów, nawet w Google.

Uświadamianie istnienia ukrytych zasobów Web może przybierać różne formy, od prezentacji w ramach formalnych szkoleń, wykładów, indywidualnych instruktaży itp., poprzez informacje i instrukcje na stronach internetowych[19], aż do notatek i artykułów w czasopismach fachowych.

Tab. 1 Wyszukiwarki Deep Web, wyszukiwarki specjalistyczne, katalogi tematyczne (głównie naukowe)

Academic Index http://www.academicindex.net/	Multiwyszukiwarka przeszukująca tysiące baz danych i serwisów naukowych oraz informacyjnych tworzonych przez biblioteki i konsorcja.
Academic Info http://www.academicinfo.net/	Katalog tematyczny dla uczniów i studentów zawierający 25 tys. wyselekcjonowanych źródeł sieciowych.
Alacra http://www.alacrastore.com/	Multiwyszukiwarka przeszukująca ponad 200 mln raportów firm oraz artykułów z czasopism biznesowych. Wyszukiwanie bezpłatne, dostęp do pełnych tekstów – płatny.
AllNewspapers.com http://www.allnewspapers.com/general/aboutus.htm	Katalog gazet, rozgłosni radiowych, stacji telewizyjnych I agencji prasowych z całego świata.
Bartleby.com http://www.bartleby.com/	Wyszukiwarka haseł ze słowników i encyklopedii, cytatów oraz informacji o popularnych pisarzach i poetach wraz z tekstami ich prac.
BUBL http://bubl.ac.uk/	Katalog źródeł internetowych ze wszystkich dziedzin wiedzy akademickiej tworzony przez bibliotekarzy brytyjskich. Możliwość przeglądania wg dziedzin, typów źródeł i krajów oraz wyszukiwania wg słów kluczowych.
CiteSeer – Scientific Literature Digital Library http://citeseer.ist.psu.edu/	Wyszukiwarka naukowa działająca na zasadzie indeksu cytowań. Baza CiteSeer zawiera ok. 770 tys. dokumentów.
CompletePlanet http://www.completeplanet.com/	Wyszukiwarka baz danych. Daje dostęp do ponad 70 tys. baz i serwisów wyszukiwawczych. Możliwość wyszukiwania wg słów kluczowych lub przeglądania wg kategorii tematycznych.
Digital Librarian http://www.digital-	Katalog zasobów sieciowych wg dziedzin, opracowywany na bieżąco przez

[19] Zob. http://www.lagcc.cuny.edu/library/invisibleweb/default.htm, http://www.vts.intute.ac.uk/

librarian.com/	bibliotekarzy amerykańskich.
DOAJ http://www.doaj.org/	*Directory of Open Access Journals*. Pełne teksty artykułów z ponad 2500 naukowych czasopism elektronicznych, dostępnych bezpłatnie w sieci.
Ekonomia Online http://kangur.ae.krakow.pl/Biblioteka/Ekonomia/	Serwis tematyczny, kierującym zainteresowanych ekonomią do ponad 1800 internetowych źródeł informacji. Wydawany przez Bibliotekę Główną Akademii Ekonomicznej w Krakowie. Możliwość przeglądania i wyszukiwania.
ERIC http://www.eric.ed.gov/	Centrum informacji i biblioteka cyfrowa źródeł edukacyjnych. Dostęp do ponad 1,2 mln rekordów bibliograficznych artykułów z czasopism i innych źródeł, z których wiele kieruje do pełnych tekstów publikacji.
Find Articles http://www.findarticles.com/	Wyszukiwarka milionów artykułów ze znanych czasopism – większość dostępna w pełnych tekstach.
Galaxy http://www.galaxy.com	Wyszukiwarka i katalog tematyczny źródeł naukowych ze wszystkich dziedzin wiedzy.
Geniusfind http://www.geniusfind.com/	Katalog tematyczny wyszukiwarek specjalistycznych i baz danych.
Google Scholar http://scholar.google.com	Specjalistyczna wyszukiwarka Gogle rejestrująca pełne teksty lub abstrakty prac naukowych (artykułów, raportów, książek), w tym linki do publikacji z wydawnictw komercyjnych.
GoshMe http://goshme.com/	Multiwyszukiwarka (wersja beta - obecnie wymagana bezpłatna rejestracja), która przeszukuje ponad 2500 wyszukiwarek i baz danych, z podziałem na wyszukiwarki specjalistyczne (np. Scirus, Find Articles, USPTO patents database) i standardowe (np. Google, Yahoo!). Wiele z tych drugich zawiera zasoby niedostępne dla wyszukiwarek ogólnych.
GrayLIT Network http://graylit.osti.gov/	Wyszukiwarka raportów technicznych. Przeszukuje kilka różnych baz danych jednocześnie.
HighBeam™ Research http://www.highbeam.com/	Multiwyszukiwarka kierująca pytanie do trzech obszarów sieci: Library (ponad 35 mln dokumentów, głównie artykułów z czasopism - płatnych i bezpłatnych), Web (zasoby sieciowe), Reference (słowniki, encyklopedie, tezaurusy)
HighWire Press http://highwire.stanford.edu/	Repozytorium zawartości ponad tysiąca czasopism naukowych i ponad 4 tys. artykułów ze 130 wydawnictw uniwersyteckich. Ok. 1,5 mln artykułów występuje w wersji pełnotekstowej (bezpłatnie).
INCYWINCY http://www.incywincy.com/	Multiwyszukiwarka zasobów „głębokich" i„płytkich". Wyszukuje w Open Directory Project, kilku ogólnych wyszukiwarkach i ponad milionie portali tematycznych.
Infomine http://infomine.ucr.edu/	Kolekcja źródeł dla środowisk akademickich, tworzona przez bibliotekarzy. Rejestruje bazy danych, czasopisma elektroniczne, książki elektroniczne, biuletyny, listy dyskusyjne, katalogi, artykuły, wykazy naukowców itp.
Infoplease.com http://www.infoplease.com/	Wyszukiwarka baz danych *deep web*. W rezultatach otrzymujemy informacje z encyklopedii, słowników i innych źródeł.
Internet Archive http://www.archive.org/index.php	Biblioteka internetowa dla naukowców, głównie historyków, oferująca dostęp do kolekcji historycznych w postaci cyfrowej.
Internet Public Library (IPL) http://www.ipl.org/	Publiczna "biblioteka internetowa" tworzona przez University of Michigan School of Information. Kolekcja linków do źródeł internetowych, pogrupowanych w kilku kategoriach. Zawiera też teksty ok. 20 tys. książek.
Intute http://www.intute.ac.uk/	Wielodziedzinowy *subject gateway* tworzony przez uniwersytety brytyjskie, dający dostęp do wyselekcjonowanych i zrecenzowanych przez specjalistów źródeł sieciowych, przeznaczonych dla nauki i edukacji. Baza danych zawiera ok. 115 tys. rekordów.
Invisible Web Directory http://www.invisible-web.net/	Przebudowywany obecnie katalog zasobów ukrytych Web. Twórcy zapewniają, że wkrótce będzie dostępny.
Librarians' Internet Index (LII) http://lii.org/	Serwis tworzony przez bibliotekarzy amerykańskich dla użytkowników bibliotek publicznych. Rejestruje ponad 20 tys. źródeł internetowych w układzie przedmiotowym. Wyszukiwanie ułatwia interfejs do zadawania pytań według różnych kryteriów formalnych i rzeczowych.
Library Spot http://www.libraryspot.com/	Centrum sieciowych zasobów bibliotecznych i informacyjnych dla nauczycieli, uczniów, bibliotekarzy i innych zainteresowanych. Źródła są selekcjonowane i oceniane przez zespół redaktorów.
Live Search (MSN) – Academic http://search.live.com/ albo http://academic.live.com/	Usługa Live Academic Serach wyszukiwarki MSN, umożliwiająca wyszukiwanie publikacji naukowych z czasopism i repozytoriów OA. W rezultacie otrzymujemy opisy bibliograficzne ze streszczeniami. Academic Search współpracuje z bibliotekami i instytucjami, aby umożliwić im dostęp do pełnych tekstów publikacji, których są subskrybentami.
MagPortal.com http://magportal.com/	Wyszukiwarka i katalog tematyczny artykułów z gazet, tygodników i czasopism popularno-naukowych.
OAIster http://oaister.umdl.umich.edu/o/oaister/	Projekt University of Michigan, którego celem jest połączenie różnych kolekcji cyfrowych, trudno dostępnych dla wyszukiwarek. Zasoby zawierają obecnie

	ponad 11 mln rekordów z 766 instytucji (w tym kilku polskich). W bazie danych zgromadzone są zasoby bibliotek cyfrowych różnych instytucji, repozytoriów instytucjonalnych i czasopisma elektronicznych.
On-Line Books Page http://digital.library.upenn.edu/books/	Wykaz i wyszukiwarka ponad 25 tys. książek dostępnych w sieci bezpłatnie.
OpenDOAR directory of open access repositories http://www.opendoar.org/	Kontrolowany katalog akademickich repozytoriów Open Access. Możliwość przeglądaniawg kontynentów, wyszukiwania repozytorium wg różnych kryteriów i przeszukiwania zasobów wszystkich repozytoriów.
Picsearch http://www.picsearch.com/	Wyszukiwarka grafik (ponad 1,7 mld obrazków w bazie)
Przewodnik WiMBP im. J. Piłsudskiego w Łodzi http://www.wimbp.lodz.pl/informacja/	Przewodnik z wyszukiwarką po źródłach informacyjnych dostępnych w Internecie, przydatnych w codziennej pracy informacyjnej. Przeglądanie wg UKD, w poddziałach znajdują się odnośniki z adnotacjami do tematycznych portali, instytucji oraz baz w Polsce i na świecie.
ResourceShelf http://www.resourceshelf.com/	Serwis redagowany pod kierunkiem G. Price'a, informujący codziennie o ciekawych źródłach sieciowych.
ROAR Registry of Open Access Repositories http://roar.eprints.org/index.php	Archiwum ok. 800 repozytoriów OA (w tym też bibliotek cyfrowych). Możliwość przeglądania wg krajów, oprogramowania, typów zawartości oraz przeszukiwania treści repozytoriów wg słów z publikacji, autorów itp.
Science.gov http://www.science.gov/	Katalog i wyszukiwarka zasobów naukowych tworzonych przez organizacje rządowe USA.
ScienceResearch.com http://www.scienceresearch.com	Portal firmy Deep Web Technologies umożliwiający dostęp do wielu naukowych czasopism i baz danych (częściowo płatnych).
Scirus http://www.scirus.com	Wyszukiwarka naukowa umożliwiająca dostęp do ponad 300 mln stron internetowych, w tym: 1) odnośników do witryn naukowych, uczelnianych, technicznych i medycznych, 2) raportów, artykułów recenzowanych, opisów patentowych, preprintów i czasopism.
Scout Archives http://scout.wisc.edu/Archives/index.php	Kontrolowany pod względem jakości katalog zasobów sieciowych i list dyskusyjnych, zawierający 23 174 rekordy ze streszczeniami. Możliwość wyszukiwania oraz przeglądania wg klasyfikacji Biblioteki Kongresu.
Singingfish http://www.singingfish.com	Wyszukiwarka plików audio i wideo.
TechXtra http://www.techxtra.ac.uk/	Bezpłatny serwis przeszukujący jednocześnie 30 baz danych i innych serwisów serwisów zakresu inżynierii, matematyki i informatyki. Ułatwia dostęp m.in. do książek, artykułów, stron WWW, ogłoszeń, raportów technicznych, technicznych-printów, rozpraw naukowych. Sporo materiałów należy do *deep web* i nie może być wyszukana przez Gogle.
Trovando http://www.trovando.it	Multiwyszukiwarka przeszukująca specjalistyczne narzędzia w zależności od typów plików czy rodzaju informacji.
Turbo10 Search Engine http://turbo10.com/	Multiwyszukiwarka przeszukująca standardowo indeksy: about.com, ask.com, dmoz.org, mirago.co.uk, search.msn.com, webfinder.com, wisenut.com, yahoo.com, yell.com. Możliwość dodania dowolnej wyszukiwarki (w tym *deep web*) z listy ok. 800, w tym Scirus i innych – opcja *Edit My Collections*.
Vascoda http://www.vascoda.de/	Portal typu *subject gateway* tworzony przez biblioteki niemieckie, oferujący dostęp do informacji naukowej z różnych dziedzin, w tym zasobów ukrytych.
Weblens - The Invisible Web http://www.weblens.org/invisible.html	"Brama" do tysięcy narzędzi wyszukiwawczych i źródeł, w tym: wyszukiwarek, katalogów tematycznych, multiwyszukiwarek, wyszukiwarek plików dźwiękowych i graficznych, wyszukiwarek ludzi i firm, informatorów, baz danych miejsc pracy, wyszukiwarek naukowych i innych.
Wikimedia Commons http://commons.wikimedia.org	Kolekcja plików graficznych do bezpłatnego wykorzystania. Ja piszą twórcy „repozytorium 1 320 274 multimediów, które każdy może rozwijać"

Zbiorcze wykazy baz danych i wyszukiwarek specjalistycznych oraz zasobów deep web:

1. About.com - Find Out More About The Deep Web - Deep Web Search http://websearch.about.com/od/invisibleweb/
2. Deep Web Research – M. Zillman`s Blog http://www.deepwebresearch.info/
3. Derfert-Wolf L.: Serwisy tematyczne o kontrolowanej jakości w Internecie – subject gateways. Wykaz. http://ebib.oss.wroc.pl/2004/57/wykaz.php
4. Gruchawka, S. R. Using the Deep Web : A How-to Guide for IT Professionals. http://www.techdeepweb.com
5. Kay B. Discovering The Invisible Web. http://lakenet.org/net_ref/manuals/invisible.html
6. Lackie R. J. Those Dark Hiding Places: The Invisible Web Revealed. Rider University. http://www.robertlackie.com/invisible/index.html
7. Pinakes - Heriot-Watt University Library http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html

8. Price G. Direct search. http://www.freepint.com/gary/direct.htm
9. Research Beyond Google: 119 Authoritative, Invisible, and Comprehensive Resources. http://oedb.org/library/college-basics/research-beyond-google
10. Tool Kit for the Expert Web Searcher, LITA American Library Association, http://wikis.ala.org/lita/index.php/Tool_Kit_for_the_Expert_Web_Searcher
11. Zillman, M. P. Academic and Scholar Search Engines and Sources – An Internet MiniGuide Annotated Link Compilation. http://whitepapers.virtualprivatelibrary.net/Scholar.pdf
12. Zillman, M. P. Deep Web Research Research 2007. http://www.llrx.com/features/deepweb2007.htm

Literatura

1. Bergman M. K.: *The Deep Web: surfacing hidden value* [online]. The Journal of Electronic Publishing, vol. 7, issue 1 [dostęp 22.12.2006]. Dostępny w WWW: http://www.press.umich.edu/jep/07-01/bergman.html.
2. Blakeman, K. Specialist tools for tackling the hidden web. W: INFORUM 2006: 12th Conference of Professional Information Resources. [online] Prague May 23-25, 2006 [dostęp 4.04.2007]. Dostępny w WWW: http://www.inforum.cz/inforum2006/pdf/Blakeman_Karen.pdf.
3. Cisek, S., Sapa, R. *Komunikacja naukowa w Internecie – mity i rzeczywistość*. Preprint [online], 2006 [dostęp 4.04.2007]. Dostępny w WWW: http://eprints.rclis.org/archive/00009035/
4. Cohen, L. *The Deep Web*. [online] University at Albany, 2006 [dostęp 4.04.2007]. Dostępny w WWW: http://www.internettutorials.net/deepweb.html.
5. Derfert-Wolf, L. *Serwisy tematyczne o kontrolowanej jakości w Internecie – subject gateways* [online]. Biuletyn EBIB 2004/6 [dostęp 4.04.2007]. Dostępny w WWW: http://ebib.oss.wroc.pl/2004/57/derfert.php.
6. Devine, J., Egger-Sider, F. *Beyond Google: The Invisible Web in the Academic Library*. The Journal of Academic Librarianship 2004, Vol. 30 No. 4, s. 265-269.
7. Gulli, A., Signorini, A. *The Indexable Web is More than 11.5 billion pages*. [online] 2005 [dostęp 4.04.2007]. Dostępny w WWW: http://www.cs.uiowa.edu/~asignori/web-size/size-indexable-web.pdf.
8. Heliński, M., Mazurek, C., Parkoła, T., Werla, M. *Biblioteka cyfrowa jako otwarte, internetowe repozytorium publikacji.* W:III konferencja: Internet w bibliotekach. Zasoby elektroniczne: podaż i popyt. [online] Wrocław, 12-14.12.2005 r. [dostęp 4.04.2007]. Dostępny w WWW: http://www.ebib.info/publikacje/matkonf/iwb3/artykul.php?f
9. Lewandowski, D., Mayr, P. *Exploring the Academic Invisible Web*. [online] Preprint, 2006 [dostęp 4.04.2007]. Dostępny w WWW: http://www.durchdenken.de/lewandowski/doc/LHT_Preprint.pdf.
10. Łamek A. *Ukryty Internet*. Magazyn Internet 2002 nr 7. s. 58-60.
11. Pamuła-Cieślak N. *Typologia zasobów Ukrytego Internetu*. Przegląd Biblioteczny 2006 z. 2, s. 153-164.
12. Pamuła-Cieślak N. *Zjawisko Ukrytego Internetu – rola bibliotek w upowszechnianiu jego zasobów* [online] 2006 [dostęp 4.04.2007]. Dostępny w WWW: http://bg.p.lodz.pl/konferencja2006/materialy/Natalia_Pamula.pdf.
13. Sherman, C. *Search for the Invisible Web*. [online] Guardian Unlimited 6.9.2001 [dostęp 4.04.2007]. Dostępny w WWW: http://www.guardian.co.uk/online/story/0,3605,547140,00.html.
14. Sherman, C., Price, G. *The Invisible Web: Uncovering Information Sources Search Engines Can't See*. Medford, NJ: Information Today, 2001.
15. Szumilas, D. *Kop głębiej! Google to nie wszystko*. Magazyn Internet - sierpień 2005, s. 60-63.
16. Zimnicki M. *Kto szuka nie błądzi*. Magazyn Internet, styczeń 2006, s. 24-31.