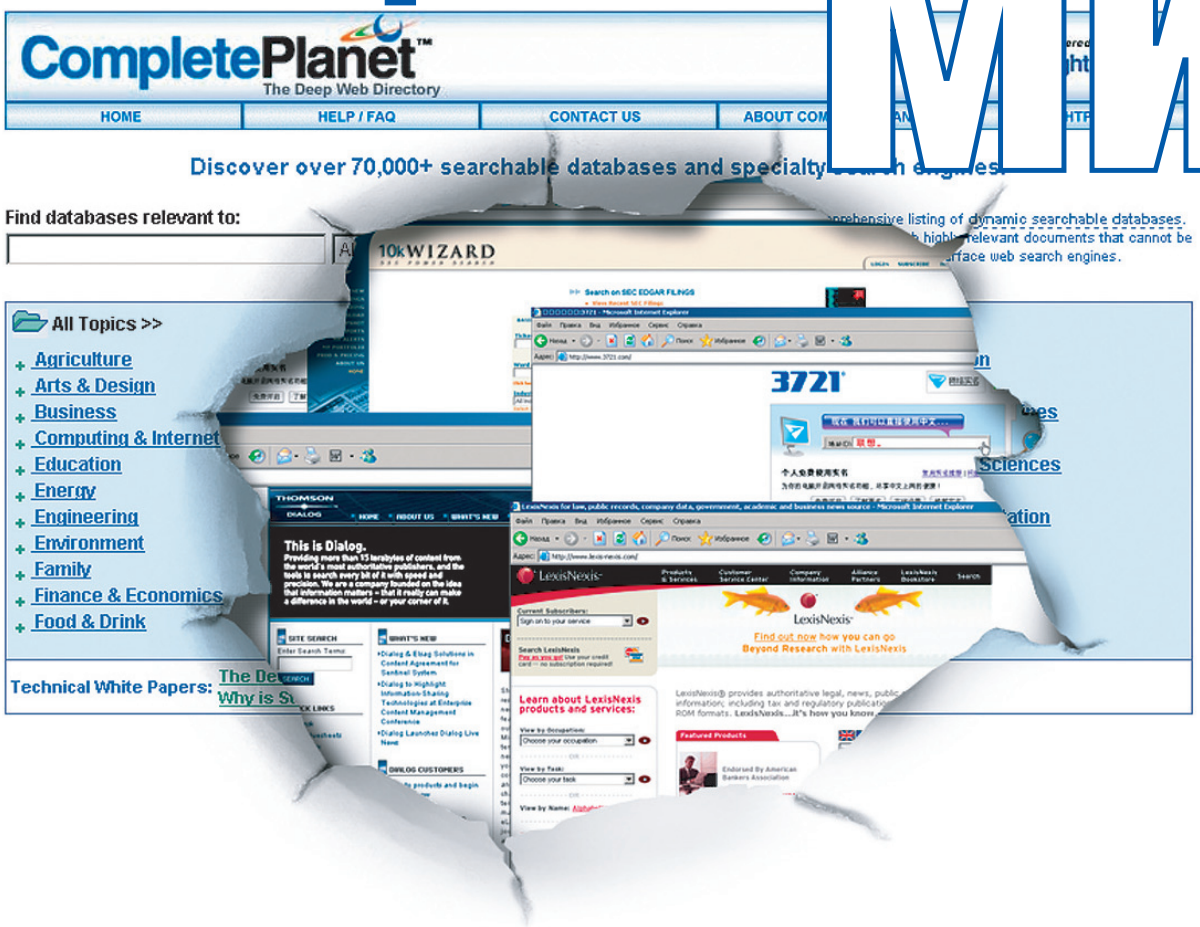
ДОПОЛНИТЕЛЬНЫЕ ВОЗМОЖНОСТИ ПОИСКА В СЕТИ

# Затерянный Мир




**Информация, доступная нам через «поисковики», — лишь малая часть того, что на самом деле находится в Сети. Чтобы открыть для себя весь необъятный мир вэба, нужно знать, как искать**

*Дмитрий Ландэ, dwl@visti.net*

Чаще всего, когда нам необходима какая-либо информация в Интернете, мы прибегаем к помощи популярных информационно-поисковых систем (ИПС), таких как Google, Yahoo!, AltaVista или «Яндекс». Однако, как оказалось, кроме видимой для «поисковиков» части вэб-пространства существует огромное количество страниц, которые ими не охватываются.

Подобные ресурсы имеют собственное название — «скрытый» вэб (deep web), которое обозначает источники, недоступные для обычных поисковых систем. Однако к таким вэб-страничкам все же можно добраться и найти на них много полезной информации.

## Подводная часть айсберга

В 2000 году американская компания BrightPlanet с помощью программы LexiBot осуществила сканирование в Сети некоторых динамических вэб-страниц, формируемых из баз данных.

Результаты были ошеломляющими — неопознанных ресурсов в Интернете оказалось в сотни раз больше, чем доступных сегодня традиционным информационно-поисковым системам. То есть 10 миллиардов вэб-страниц — это лишь видимая крупица Глобальной паутины. В результате исследований также выявилось немало интересных особенностей «скрытых» вэб-страниц. Например, оказалось, что они в среднем на 27 % компактней страниц из видимой части Интернета.

Невидимыми ресурсы для ИПС оказались по причине алгоритмов работы популярных роботов-индексаторов. Такие программы, как правило, посещают вэб-страницы по известным заранее адресам, анализируют их содержание и выделяют гиперссылки, идущие от них. Обычно, обработав текущую страницу, выделив ключевые слова и некоторые поля, робот переходит по адресам, найденным на ней. Затем система опять сканирует последующие страницы и выделяет новые адреса. Однако как только робот определяет, что он обращается к динамической странице, его работа приостанавлива-

ется. Ведь для получения осмысленного ответа из баз данных требуется соответствующий запрос, а большинству из роботов чужды элементы интеллекта (даже искусственного).

То есть «скрытый» вэб формируется в первую очередь из содержимого онлайновых баз данных. Сюда следует добавить и быстро обновляемые ресурсы, так называемые динамические. К ним относятся новости, конференции, онлайновые журналы и др. Конечно, есть и «острова» Глобальной паутины, на которые не ведут никакие гиперссылки и от которых гиперссылки не исходят. Защищенные паролями вэб-сайты также попадают в категорию «скрытого» Интернета.

О материалах этих сайтов большинство пользователей никогда не узнают с помощью стандартных поисковых систем. Однако относительное количество таких ресурсов невелико. По мнению ученых, среди крупнейших сайтов «скрытого» вэба платными являются только 10 % ресурсов.

## Многоликость ресурсов

В свое время исследователи BrightPlanet определили более десятка разновидностей «скрытых» вэб-ресурсов, относящихся к классу онлайновых баз данных. В списке оказались как традиционные базы данных (патенты, медицина и финансы), так и публичные ресурсы — объявления о поиске работы, чаты, библиотеки, справочники. Ученые также причислили к «скрытым» ресурсам и специализированные поисковые системы, которые обслуживают определенные отрасли или рынки. При этом базы данных таких ИПС не включаются в каталоги глобальных поисковых служб.

К «невидимой» части Сети также относятся многочисленные системы интерактивного взаимодействия с пользователями — помощи, консультации, обучения, требующие участия людей для формирования динамических ответов от серверов. В ней также находится и закрытая (полностью или частично) информация, доступная пользователям Сети только с определенных адресов, групп адресов, а иногда городов и стран.

Например, для нашего пользователя наверняка «скрытой» можно признать большую часть гигантского китайского сегмента Интернета. Так, малоизвестный в Европе и Америке китайский поисковый портал Baidu (*www.baidu.com*) в 2004 году опередил Google по объему трафика, став четвертым в мире вэб-ресурсом по этому показателю. Другая китайская поисковая система 3721.com (*www.3721.com*) заняла седьмое место в списке самых посещаемых ИПС.

## Сайты-невидимки

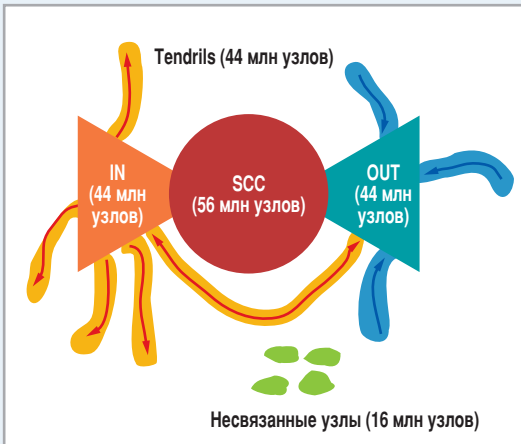
Как это ни парадоксально, но к скрытой части Сети порой относятся и крупнейшие мировые службы информационного поиска. Например, портал Dialog (*www.dialog.com*), который сегодня принадлежит корпорации Thomson (США) — одной из всемирных лидеров в предоставлении информационных сервисов в области бизнеса, науки, финансов, законодательства и др. На Dialog находится более 900 баз данных, из которых пользователи ежемесячно просматривают порядка 17 миллионов страниц документов с различной информацией. При этом Thomson определяют эти ресурсы как часть «скрытого» вэба, заявляя, что они содержат полезной (не дублирующейся) информации в 500 раз больше, чем доступно с помощью традиционных информационно-поисковых систем.

В «скрытом» Интернете также существует множество альтернатив базам данных типа Dialog. Среди них, например, сайт *www.10kwizard.com*, предлагающий доступ к полным текстам корпоративных документов, хранящихся в Комиссии США по ценным бумагам и биржам. Другой ресурс — Educator's Reference Desk (*www.askeric.org*) содержит свыше двух тысяч учебных планов, несколько тысяч ссылок на образовательные документы. С этого сайта обеспечивается доступ к базе данных ERIC — крупнейшему источнику информации по проблемам образования, а также к полнотекстовым дайджестам, составляемым

## Топология Паутины

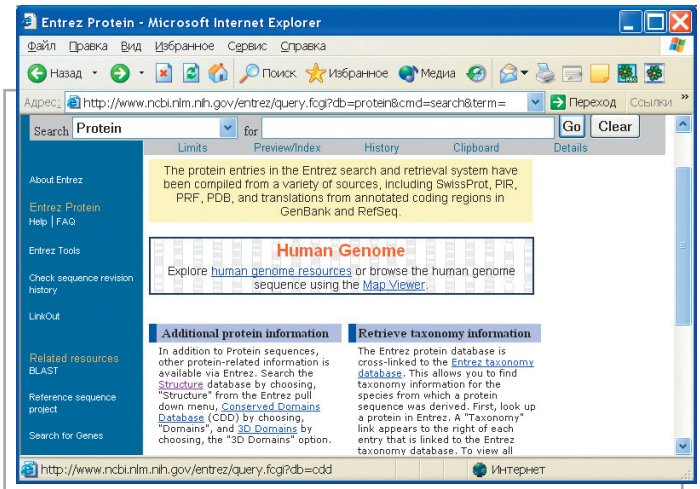
В ноябре 1999 года Андрей Бродер (Andrei Broder) совместно с другими учеными из компаний AltaVista, IBM и Compaq математически описали карту ресурсов и гиперсвязей Интернета. Исследователи опровергли расхожее мнение, будто Интернет — это единое густое пространство. Проследив с помощью поискового механизма AltaVista свыше 200 млн вэб-страниц и несколько миллиардов ссылок, ученые построили ориентированный граф с топологией «галстук-бабочка» (Bow Tie), которая, по их мнению, соответствует структуре вэб-пространства. У ученых получилась модель связности, охватывающая более 90 % исследованных вэб-страниц в Сети:

- центральное ядро (SCC) составляют ресурсы, взаимосвязанные так тесно, что, следуя гиперссылкам, из любой из них в конечном счете можно попасть на другую страницу. Они занимают порядка 28 % вэб-страниц Интернета;
- отправные вэб-страницы (IN) содержат гиперссылки, которые в конечном счете ведут к ядру, но из ядра к ним попасть нельзя.Таких страниц 22 % в Сети;
- оконечные страницы (OUT) составляют 22 % всех ресурсов. На них можно попасть по ссылкам из ядра, но нельзя вернуться обратно;
- отростки (Tendrils) — области вэб-страниц, которые оказались полностью изолированы от центрального ядра, но на них можно попасть из областей IN и OUT. Эти ресурсы составляют 22 %;
- несвязанные узлы (всего 6 %) представляют собой островки без связи с другими вэб-страницами.




Структура вэб-пространства, по мнению ученых, представляет собой топологию «галстук-бабочка» (Bow Tie)

## Скрытые страницы Интернета




Поисковая система NCBI охватывает многочисленные базы данных по медицине и естествознанию

экспертами. Австралийский ресурс Nuclear Explosions Database (*www.ga.gov.au/oracle/nukexp_query.html*) содержит базу данных по географии. Для работы с системой достаточно перейти в режим Online Tools, после чего будет представлен список баз данных и карт.

Вэб-портал PubMed (*www.ncbi.nlm.nih.gov/entrez/query.fcgi*) обеспечивает доступ к более чем 14 миллионов ссылок на материалы американской Национальной библиотеки по медицине (National Library of Medicine), включая ссылки на полные тексты статей и информационные ресурсы. С портала обеспечивается также доступ к глобальной поисковой системе NCBI (*www.ncbi.nlm.nih.gov*), охватывающей базы данных по естествознанию и медицине.

Приведем пример еще одной, весьма интересной «скрытой» базы данных в Сети. Так, корпорация ChoicePoint недавно предоставила сервис AutoTrackXP (*www.autotrackxp.com*), вошедший в список двадцати крупнейших «скрытых» сайтов мира (по рейтингу BrightPlanet). AutoTrackXP представляет собой базу данных объемом 30 терабайт, охватывающую практически все аспекты гражданской жизни США. Эта база данных содержит информацию практически о каждом гражданине США. Например, чтобы определить, не завладел ли человек чужими документами, на системе организован платный сервис ProCheck, позволяющий сопоставить информацию из различных государственных каталогов.

Парадоксально, но как ресурсы «скрытого» вэба можно рассматривать и некоторые архивы общедоступного вэб-пространства. Например, такой архив Internet Archive компании Alexa (*www.alexa.com*) содержит базы данных объемом более 500 терабайт. Новые страницы в настоящее время попадают в хранилище со скоростью 1 терабайт в день.

Технология хранилища Alexa включает ряд современных средств управления гигантским контентом. Например, с ее помощью выполняется кластеризация вэб-ресурсов, то есть формирование коллекций документов, близких по тематикам. Особый интерес у пользователей Alexa вызывает сервис «Машина времени» (Wayback Machine). Он позволяет восстанавливать документы, некогда опубликованные в Интернете, но впоследствии удаленные.

Традиционная поисковая система чаще всего может назвать адрес базы данных, но не скажет, какие документы конкретно содержатся в ней. Типичный пример — информационно-поисковые системы по украинскому (*www.rada.gov.ua*) и российскому (*www.kodeks.ru*) законодательству. Тысячи документов из их баз данных становятся доступны только после входа в систему. При этом роботы стандартных ИПС не в состоянии проиндексировать контент подобных баз данных.

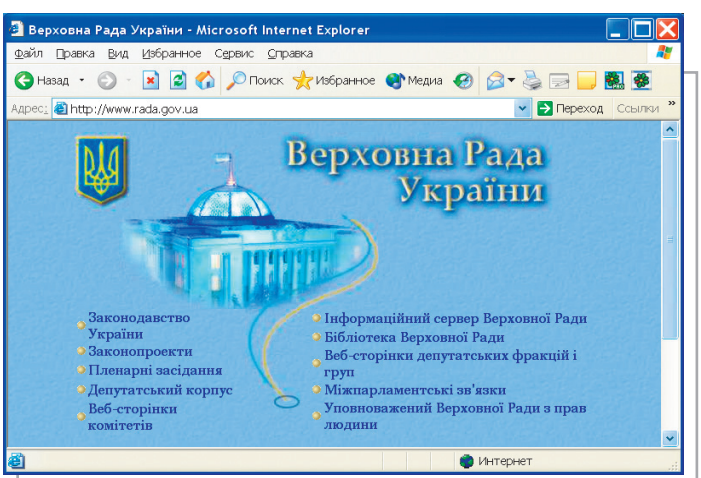
### Сталкеры Сети

«Скрытый» вэб представляет собой гигантское хранилище документов, звуков, изображений, фильмов и др. Безусловно, если большая часть этой информации не доступна традиционным поисковым системам, то существует потребность в специальных инструментах для поиска «скрытого» контента.

Так, для поиска в «скрытой» Сети (а именно в том ее сегменте, который составляют базы данных) сегодня уже существуют некоторые специализированные ресурсы. Среди них, например, система Invisible Web (*www.invisible-web.net*) компании IntelliSeek. Сайт включает каталоги баз данных, большинство из которых не проиндексированы известными поисковыми системами. При введении запроса этот механизм выдает ссылки на ресурсы, с помощью которых поиск необходимой информации станет наиболее оптимальным. На этом вэб-сайте также собраны коллекции ссылок на различные базы данных, среди которых содержится немало уникальных ресурсов, например, сборник выступлений и докладов известных политиков и бизнесменов.

Известным навигатором в «скрытом» Интернете является и сайт CompletePlanet (*www.completeplanet.com*) компании BrightPlanet. Этот ресурс является крупнейшим каталогом, насчитывающим свыше 100 тыс. ссылок. Специальный метапоисковый пакет DeepQuery-Manager этой же компании обеспечивает поиск по 55 тыс. «скрытых» вэб-ресурсов.

Вэб-сайт Direct Search (*www.freepint.com/gary/direct.htm*) также




Для традиционных «поисковиков» скрыты и базы данных с украинского портала «Верховна Рада України»

## Традиционные ИПС — шаг вперед

■ Информация, представленная в форматах, отличных от HTML, для многих известных поисковых систем до последнего момента оказывалась недоступной. Однако сегодня ситуация меняется в лучшую сторону. Например, популярная система Google (*www.google.com*) уже обеспечивает поиск в документах, представленных в форматах MS PowerPoint, DOC, RTF, PostScript и PDF, а также осуществляет преобразование этих файлов в текстовый формат. При этом поиск документов в разнообразных форматах доступен в этой системе как из режима расширенного поиска (Advanced Search), так и из обычного поиска — достаточно использовать в запросе команду filetype: (например, для файлов PDF это будет filetype:pdf). Другая известная служба Yahoo! (*www.yahoo.com*) также обеспечивает выдачу текстовых копий документов, размещенных в форматах Word, Excel, PowerPoint, PDF, RSS/XML-фидов (новостных лент и блогов — «живых журналов»). В свою очередь, специализированная система Gigablast (*www.gigablast.com*) предназначена исключительно для поиска по документам в форматах Word, Excel и PDF. Она выдает по запросу кэшированные (архивные) копии документов в исходных форматах, при этом обеспечивает поиск и выдачу копий документов, которые были размещены в Сети, но затем, возможно, удалены.

обеспечивает поиск в базах данных «скрытого» вэба. На сайте содержатся ссылки на лучшие ресурсы ценовой (MySimon.com) и финансовой (FinancialFind.com) информации, а также ссылки на ресурсы научно-популярных журналов и научных баз данных (Biolinks.com).

В Интернете есть и другие сайты-навигаторы, а также специализированные программы поиска. Например, Infomine Multiple Database Search (*http://infomine.ucr.edu/search.phtml*) — поисковая система по университетским архивам, библиотекам и книгам; Bubl Link (*www.bubl.ac.uk/link*) — каталог информационных сайтов, посвященных различным областям человеческой деятельности; Amazon.com (*www.amazon.com*) — полнотекстовый поиск по содержанию всех книг.

Стоит отметить, что особенностью большинства ресурсов-невидимок является их узкая специализация. Поэтому роботы поисковых систем для «скрытого» вэба включают уникальные для каждого такого ресурса модули доступа к данным.

### Вэб сегодня и завтра

Ввиду роста количества вэб-сайтов, использующих в своей работе информационные базы данных и различные динамические системы управления контентом, «скрытый» сегмент Интернета растет очень интенсивно. При этом все меньшая часть информационных ресурсов становится доступной пользователям посредством традиционных поисковых механизмов.

Оказалось, что спасти ситуацию могут новые возможности унификации данных в Интернете. Одним из первых проектов консорциума W3C, занимающимся развитием Глобальной сети в этой области, стал «Семантический Вэб». Его основная идея заключается в организации данных, которая позволила бы вэб-серверам (с программами разных производителей) эффективно их использовать. В рамках проекта были разработаны спецификации метаязыка XML, предусматривающие разделение средств визуализации и смыслового содержания. На основе XML удается создавать различные форматы, специально предназначенные для организации коммуникации как между персональными устройствами, так и между серверами.

Так, например, для решения задачи интеграции новостной информации было создано несколько форматов описания данных на основе XML. Самый распространенный формат получил название RSS (см. ЧИП 7/2004, с. 82). Сегодня экспорт данных в формате RSS осуществляют крупнейшие порталы, включая CNN (*www.cnn.com*), BBC News (*www.bbc.co.uk*), CNet News (*www.cnet.com*), MSNBC (*www.msnbc.com*), The Register (*www.theregister.com*), Wired News (*www.wired.com*) и др. Аналитики отмечают, что только в 2004 году пользователи Интернета по-настоящему открыли для себя возможности технологии RSS.

Сегодня для работы с данными в формате RSS создаются новые программы, сайты и поисковые системы, которые все более востребованы пользователями. Возможно, подобные технологии и программы сумеют вскоре приоткрыть завесу над гигантской частью «скрытого» вэба.

**INFO**

Еще о скрытых вэб-ресурсах
*http://links.chip.ua/web_04*