Программы для записи на диск Веб-сайтов (Offline Browsers)

Материал подготовил Григорий Наумовец <gri@aiha.kiev.ua>.

Зачем нужны специальные программы для записи на диск Веб-сайтов?

Если создать на диске копию Веб-сайта, в дальнейшем его можно будет просматривать и использовать имеющиеся файлы без подключения компьютера к сети Интернет. Однако Веб-сайт может состоять из большого количества файлов, размещенных в различных каталогах. Для того, чтобы в локальной копии Веб-сайта работали все ссылки и связи, нужно не просто записать на диск все необходимые файлы, но и сохранить структуру их расположения, т.е. поместить каждый файл в соответствующий каталог. Сделать это "вручную" с помощью Веб-браузера – работа долгая и кропотливая, а при большом числе файлов – практически невыполнимая. Поэтому и были разработаны специальные программы, предназначенные для автоматической записи на диск Веб-сайтов.

Internet Explorer v.5 и 6: Возможность "полного" сохранения Веб-страницы на диске

Internet Explorer (IE) v.5 и 6, в отличие от 4-х версий Internet Explorer и Netscape Navigator, может сохранять Веб-страницы на диск по-разному -- "полностью" либо "только HTML". Если открыть какой-нибудь HTML-документ с помощью IE и в меню *File -> Save As* указать *Save as type: Web page, HTML only*, то на диск будет записан только сам HTML-документ. Если же указать *Save as type: Web page, complete*, то будет сохранен не только HTML-файл, но и те графические файлы, на которые в нем есть ссылки. Эти "картинки" будут записаны в специально создаваемую вспомогательную папку (например, если документ называется testdocument.html, то папка будет называться testdocument_files), и ссылки на них внутри документа будут переписаны таким образом, чтобы они указывали именно на эту папку. (Поэтому файл, сохраненный на диске, не будет точной копией файла, загруженного с сервера). Кроме того, по умолчанию IE предлагает сохранить HTML-файл не под исходным именем, а под именем, соответствующим его заголовку. Если сохранять на диск "полностью" другие HTML-документы с того же сайта, каждый раз будет создаваться новая вспомогательная папка, в которую будут записаны все графические файлы, нужные для показа данного документа (если на одну и ту же картинку есть ссылка в нескольких документах, ее копия каждый раз будет снова записана в соответствующую папку). Ссылки с одного документа на другой будут работать только в том случае, если их редактировать "вручную". Таким образом, эта функция IE удобна для сохранения на диск отдельно взятой Веб-страницы вместе с картинками, но неэффективна для сохранения многих связанных между собой файлов.

Если в меню IE *File -> Save As* указать *Save as type: Web archive, single file*, то Веб-страница вместе с картинками будет сохранена в одном файле с расширением .mht. По формату этот файл напоминает почтовое сообщение, в теле которого пересылается сам HTML-документ, а картинки прилагаются в MIME-закодированном виде. Файл типа *.mht можно открыть в IE или Netscape Navigator и увидеть Веб-страницу вместе с картинками.

Internet Explorer v.5 и 6: просмотр файлов из кэш-памяти в режиме offline

Если в IE нажать кнопку *History*, откроется окошко, в котором будет показан список сайтов, просмотренных в последнее время с помощью данного браузера. Если через меню *File -> Work Offline* включить режим *Offline*, можно попробовать просмотреть некоторые из этих сайтов без подключения к Интернету (при этом IE использует ранее загруженные файлы, хранящиеся в папке "*Temporary Internet Files*". Посмотреть, где находится эта папка, переместить её в другое место или просмотреть сами файлы можно через меню Internet Explorer'a *Tools -> Internet Options -> General -> Temporary Internet Files -> Settings*). Те сайты, которые нельзя просмотреть без подключения к Интернету, показываются бледным шрифтом.

Для преобразования набора файлов, хранящегося в кэш-памяти Internet Explorer'a или Netscape Navigator'a, в локальные копии недавно просмотренных Веб-сайтов существуют специальные программы – например, HTML Converter (http://htmlconv.da.ru).

Программы для записи на диск Веб-сайтов: какие программы можно использовать и где их взять

На Интернете можно найти много программ для записи на диск Веб-сайтов, например:

WebStripper (www.webstripper.net)
WebZip (www.spidersoft.com)
Teleport Pro (www.tenmax.com/teleport/pro/home.htm)
Offline Explorer (www.metaproducts.com)
и другие.

Кроме того, эти программы можно найти в разделе "*Internet -> Web Browsers & Tools -> Offline Browsers*" на известном сервере Tucows (www.tucows.com). Для быстрой загрузки можно выбрать одно из многочисленных зеркал сервера Tucows, в том числе:
в России: http://tucows.rinet.ru, http://tucows.online.ru
в Украине: http://tucows.uar.net

См., например, http://tucows.online.ru/offline95.html .

Как сохранить Веб-сайт на диск: данные, которые нужно сообщить программе

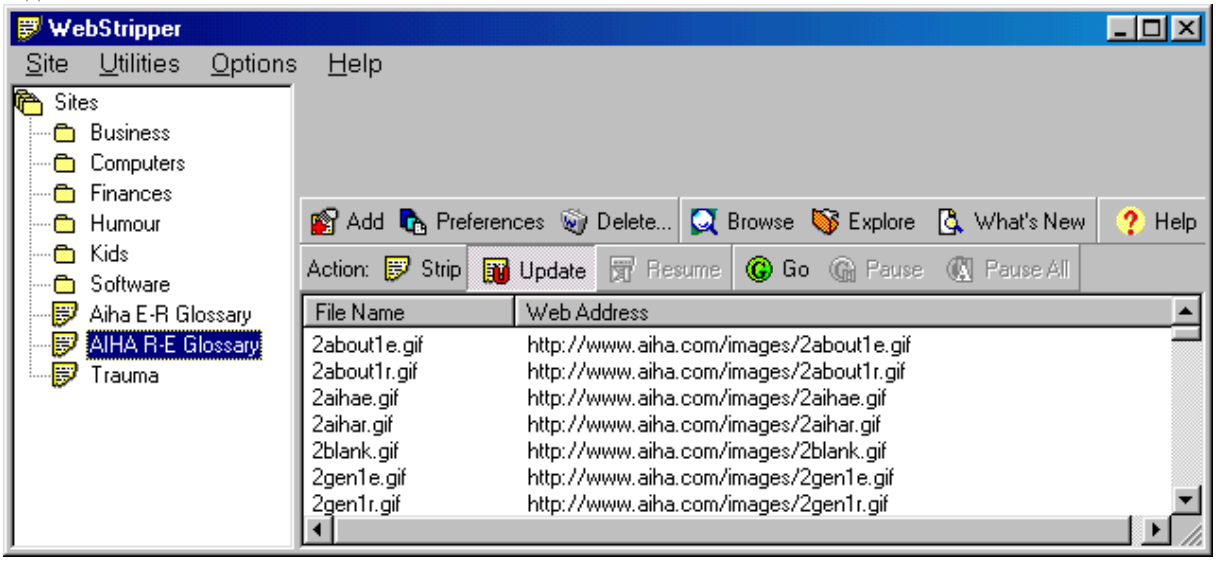
Прежде всего, программа должна "знать" URL основной Веб-страницы, с которой нужно начинать загрузку. Просмотрев исходный документ, программа может составить список всех файлов, ссылки на которые там есть. Это могут быть другие HTML-документы, графические, звуковые, текстовые и другие файлы. Одни из этих файлов могут находиться на том же Веб-сервере, что и исходный документ, другие – на других серверах. Загрузив HTML-документы, на которые ссылается исходный документ, программа найдет там ссылки на другие файлы (так сказать, "ссылки второго порядка"), в тех файлах – еще ссылки ("ссылки третьего порядка"), и так далее – теоретически, это может продолжаться до бесконечности (а реально -- пока не кончится место на диске). Поэтому необходимо задать критерии того, какие файлы загружать, а какие нет. На практике эти критерии могут выглядеть таким образом:

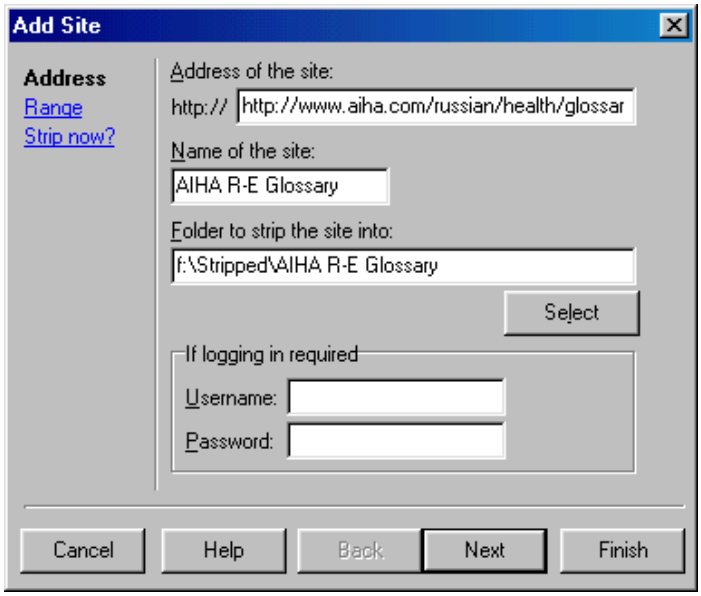
- Грузить ли только те файлы, которые находятся на данном сервере, или и с других серверов тоже?
- В пределах данного сервера, грузить файлы только из данного каталога и его субкаталогов, либо и из других каталогов тоже?
- До какой "глубины" (т.е. до ссылок какого порядка) нужно проследить "дерево" ссылок? (Эта глубина может быть разной для файлов, находящихся на данном сервере либо на других серверах).
- Грузить ли все виды файлов, или ввести какие-то ограничения (например, для звуковых файлов, или для графических, или для файлов больше определенного размера и т.п.)?

Загрузка сайтов при помощи программы WebStripper

Ввод параметров нового задания

Для того, чтобы загрузить сайт при помощи программы WebStripper, нужно сначала через меню *Site -> Add site* (либо нажатием кнопки *Add*) вызвать окошко диалога для ввода параметров нового задания.

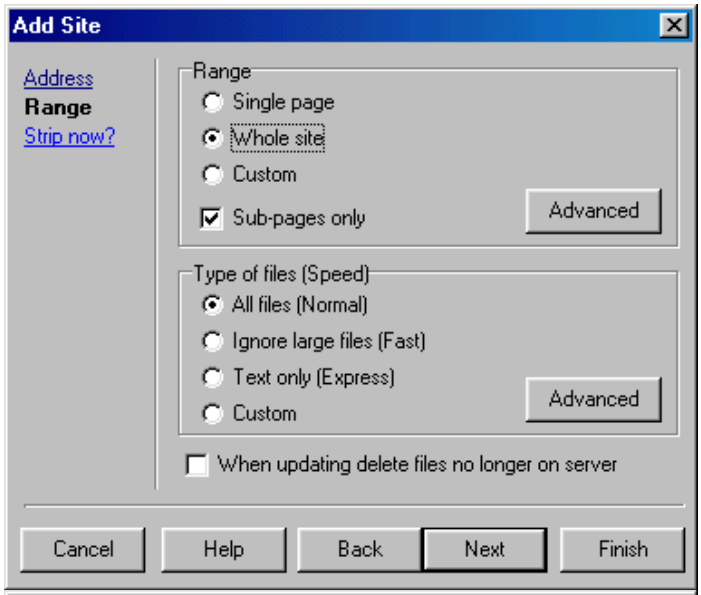




В этом окошке нужно ввести (или скопировать из браузера) начальный URL (*Address of the site*), задать некое условное имя для этого сайта (*Name of the site*) и место на диске, куда этот сайт записать (*Folder to strip the site into*).

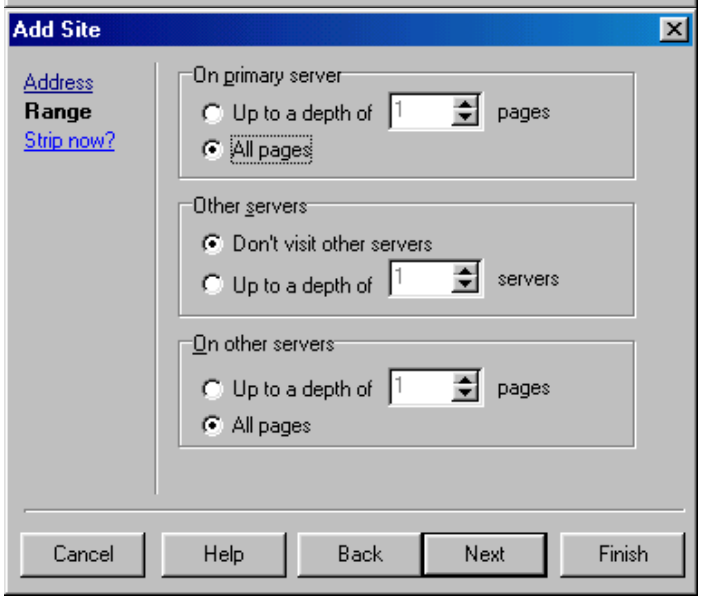
Если доступ к данному серверу защищен паролем, логин и пароль нужно указать в окошке "*If logging in is required*" (не нужно путать эти логин и пароль с теми, что используются для дозвона к провайдеру!).

Для задания критериев того, какие файлы грузить, а какие нет, нужно нажать на ссылку *Range* ("Диапазон").



В окошке Range можно указать диапазон файлов, которые нужно загружать (Single page – только данную страницу, Whole site – весь сайт, Sub-pages only – только файлы, находящиеся в данном каталоге и его субкаталогах), а также типы файлов, которые нужно либо не нужно загружать.

С помощью кнопок *Range -> Advanced* и *Type of files -> Advanced* можно перейти в окошки более детальной настройки критериев загрузки – откуда файлы грузить, а откуда нет, и какие типы файлов грузить, а какие нет.



В этом окошке (*Range -> Advanced*) можно задать, до какой глубины проходить "дерево ссылок" на исходном сервере (*On primary server*) и на других серверах (*On other servers*), а также переходить ли на другие сервера вообще (*Other servers*), и переходить ли на третьи сервера, ссылки на которые есть на других серверах и т.п. – т.е. до какой глубины проходить "дерево серверов".

Задав все критерии для загрузки файлов, можно нажать на ссылку *Strip now*, чтобы запустить процесс загрузки Веб-сайта. При этом откроется специальное окошко, в котором будет показан процесс загрузки. С помощью кнопок *Pause* и *Resume* загрузку можно приостанавливать и снова возобновлять.

Когда загрузка будет завершена, локальную копию Веб-сайта можно будет просмотреть прямо в окошке WebStripper'a при помощи встроенного браузера.

Иногда один сайт может иметь несколько имен (alias'ов) и фигурировать в различных ссылках под разными именами. WebStripper воспринимает различные имена одного и того же сайта как разные сайты.

Автоматический дозвон до Интернет-провайдера

Автодозвон до Интернет-провайдера включается и отключается через меню *Options -> Options -> Connection -> Use a dial-up Internet connection* (логин и пароль должны быть указаны в настройках самого соединения, которое должно быть сконфигурировано в Dial-up Networking). Если включена опция *Hang up when done*, соединение будет разорвано по завершении загрузки.

Настройки для работы через прокси-сервер

Если компьютер пользователя может выходить в Интернет напрямую (без прокси-сервера), параметры прокси можно не указывать.

Если же компьютер пользователя подключен к Интернету не напрямую, а через прокси-сервер, параметры этого сервера необходимо указать в настройках *Options -> Options -> Connection -> Use a proxy server*. Параметры прокси сервера можно посмотреть в настройках веб-браузера (в Internet Explorer 5: *Tools -> Options -> Connections -> LAN settings -> Proxy server*, Netscape 4.5-4.7: *Edit -> Preferences -> Advanced -> Proxies -> View*) либо узнать у провайдера или сетевого администратора.

Dial protection (вид ссылок на еще не загруженные файлы)

Если в браузере включен режим автодозвона до Интернет-провайдера, просмотр в нем записанного на диск документа, содержащего ссылки на файлы (например, картинки), которые не были сохранены на диск, либо нажатие на такие ссылки может привести к тому, что запустится программа автодозвона. Для блокирования такого автодозвона в WebStripper'е предусмотрен режим *Dial Protection*, управляемый через меню *Options -> Options -> Browsing -> Dial Protection.* Если этот режим включен (варианты *When page loads* и *When following a link*), ссылки на еще не загруженные файлы переписываются так, чтобы автодозвон не запускался – например, заменяются специальным текстом с информацией о том, что данный файл не был загружен. Если этот режим выключен (вариант *None*), ссылки на еще не загруженные файлы не переписываются.

Если нужно изменить режим *Dial Protection* для сайта, который уже был загружен, не обязательно загружать его заново. Достаточно изменить этот режим в настройках программы, а потом выделить нужный сайт и дать команду *Utilities -> Re-Parse*.

Обновление сайта (Update)

Если на сайте, который уже был загружен, добавились либо обновились некоторые файлы, не обязательно загружать его заново – достаточно войти в меню *Site -> Action ->* Update, и WebStripper автоматически проверит сайт на наличие изменений и загрузит обновившиеся файлы.

Программы WebZip и Teleport Pro

В отличие от WebStripper'а, программы WebZip и Teleport Pro являются платными, и их пробные версии имеют ограничения по сроку использования и/или по функциональным возможностям (например, в пробной версии Teleport Pro ограничено число загружаемых файлов). Однако эти программы обладают некоторыми дополнительными возможностями по "фильтрации" загружаемых файлов по сравнению с WebStripper'ом. Например, в WebZip предусмотрены специальные "запретительные" и "разрешительные" фильтры, с помощью которых можно загружать только те файлы, в URL'ах которых встречаются либо не встречаются заданные последовательности символов(меню *Download Method -> URL Filters* в свойствах задания).