HTML Title Tag Extractor

Purpose:

This application finds the content of HTML <TITLE> tags in all the HTML files of a website.

Justification:

Over a period of time, an active website, particularly one used for eCommerce, can get 'ratty'. This means that the involvement of multiple developers, 'lessons learned' by the web developers, migration from static to dynamic content, and other considerations leaves a variety of coding practices, compliance to differing coding standards, and in some cases, unfinished work populating various directories within the web.

Many web pages have no title and the user can only surmise the contents of the page. Others are badly worded, and some contain spelling, punctuation, or grammatical errors. Since this is a 'first impression' issue, periodically reviewing and correcting titles is a way to avoid embarrassment, confusion, or lost sales.

This program can, with various modifications, identify and perhaps correct repetitive and flagrant problems.

Provider Purpose and Justification:

This program is offered free, and is hereby put in the public domain by Resource Logic, Inc. The downloadable module has no encryption and can be modified by the user for any purpose whatsoever.

This program is offered as a promotional marketing tool by Resource Logic, Inc. Our purposes in making it available is first to demonstrate our technical skills, second to offer 'quick fix' education to developers that need to see a working example of particular language features, and third to find users who have either web development or Access database problems they would like us to solve. Any use of this program is at the user's own risk. Resource Logic does not warranty the program for any particular purpose, and in any event is not liable for consequential damages. The contents of this manual constitute the entire agreement, if any, between Resource Logic, Inc. and the user.

Operation

Functionality:

This application scans through a website looking for all files that are either .HTM(L)s or .PHPs and saves these names into a table. In a follow-on step, the program scans through each of these files looking for the <TITLE> and </TITLE> tags that bracket the page title. The page title is what appears in search engines when a user requests pages associated with a particular term.

The user then reviews titles on a page by page basis and enter revisions. This program does not apply revisions to the web pages (see disclaimers below). Using standard HTML editing tools like FrontPage, Dreamweaver, or GoLive, the user is able to modify page titles as appropriate.

Limitations and Assumptions:

1. Some websites use extensions other than .HTM or .PHP for their names, with .ASP being common. Many files with embedded script are barely recognizable as web pages.

2. This program does not update anything outside the database. In particular, it does not replace text in HTML files.

3. The system scans files located in a directory on an attached hard drive or network drive. It is not able to scan files that are located on web servers or are otherwise provided through HTTP or FTP. The user must first transfer these files to local directories.

4. The program was developed in Office XP and will also run in Office/2000 (Microsoft Access/2000 and XP). The user must have one of these Microsoft products installed on their computer. We highly recommend that all operating system and Office 'critical updates' be applied to any system running database applications, or any other application in which users can download scripts or executable code over the Internet.

Data Structures:

There are two tables included in the database:

StartDirectory - contains one record with one column. This contains the directory name that is the 'root' of the web. All files scanned are either in this directory or in folders that are in this directory.

PageTitles - contains the list of files and associated titles.

Forms:

There are two forms in the database:

frmScanHTML - Performs a set of functions associated with storing file names in the database and scanning for HTML tags.

frmPageTitles - Browse and maintenance form, used primarily for entering replacement titles.

There are no queries, reports, macros, or modules in the application. No menus are needed or appropriate under the circumstances: the user's default Access menus are preserved.

Scanning Files and Tags:

Clear Table deletes all records out of the PageTitles table.

Scan Directories first checks to see if a starting directory path has been entered. If not, it sets the default starting directory to the path contained in the StartDirectory table.

Note: If a pathname contains embedded spaces, please put double quotes around the entire path. For example:

X:\SampleDirectory

is fine as specified since there are no spaces or other special characters.

If there is an embedded space:

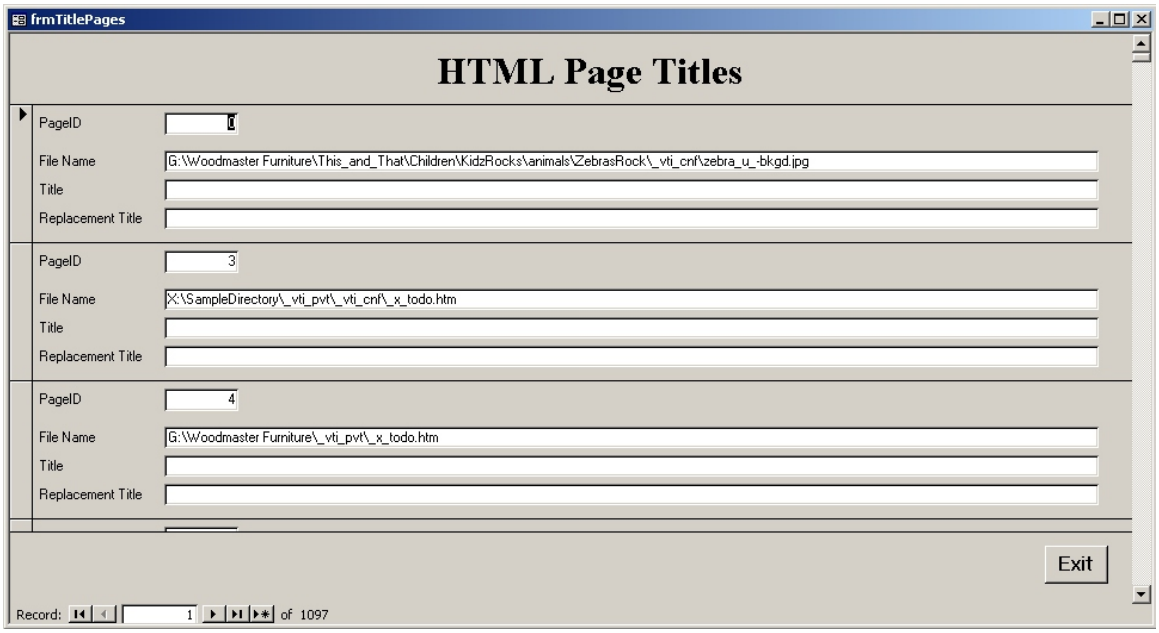
"X:\Sample Directory"

should be used in the Starting Directory Path field.

Once the directory path is in place, the program scans through all files in or subordinate to this directory. Those that contain the text '.htm' or '.php' are written to the table. There is no assurance that these files are necessarily HTML or PHP files, since someone could, for instance, name a file 'Demo.HTM.ASP' or 'Process.PHP.CGI'. Any user that has such files on their site probably knows what to do with these particular instances.

Extract Titles opens the files as they have been saved and searches for the <TITLE> tag in the text. If multiple <TITLE> tags are found, the last one is used. If there is no </TITLE> on the same text line as the <TITLE> tag, all text remaining on that line is used. Multiple line text is not included. If there is no <TITLE> tag, then the column remains empty (null). The case of the text within the tag is immaterial. <tItLe> will be scanned and recognized.

Maintaining Titles:



This form lists the sequence number of each record, the full file path, the existing title text, and a null field which the user may replace with updated text. This new field would be used to apply changes to the HTML, if that feature were offered. Since this application is 'read only' with respect to files within a web site, the 'Replacement Title' field exists only for illustration.

Notes On Code:

Programmers new to Access or VBA (Visual Basic for Applications) might want to see (or have) working examples of the following functionalities:

- ADODB
- FILESEARCH Object
- String manipulation
- DOCMD.RUNSQL

The SQL being run in this case is the MSAccess SQL (OLEDB) and would not necessarily work in SQL-Server or SQL compliant databases.

The user is not mistaken in thinking this system is a 'quickie'. This utility may exist in similar form elsewhere, including in web page development applications. In our specific situation, we will be using a derivative of this program to make 'macro' modifications to a large body of HTML files.