1. Website address is http://adiglobal.us. I have an existing spider that I use on this site. But I have been having issues with the spider the last few months. The developer has been busy on other projects and has not had enough free time to fix it. I can provide you the program if needed for reference only.

2. This website tracks foreign IP’s and block if the ping rate gets too high. You will need to limit your testing from Pakistan unless you have a PC in the USA you can test from. If not, I have a Remote PC that you can Remote Desktop into for testing. When the spider is complete, you will be required to install and execute on this Remote PC. All future updates will be handled in the same manner.

3. The only difference between the http and https side is the addition of Cost and Inventory information. The product details and images are the same. Once you successfully create a spider that collects the images and details, I will send you the login information for scraping the Cost and Inventory. You must always use the Remote PC when logged in as they geo locate the IP to my account login.

4. The spider will need to operate in two different modes. The first mode is as a web crawler, whereas it crawls through the site and saves the data to an Access database. The second mode is where is will update the Cost and Inventory using an Access database.

a. The Web Crawler mode will have the option of by Category or by Brand. When on the site, click on “shop products” which will pull up a list of categories in the left menu bar. Each of those categories has a sub-category and sub-sub-category. For example, Access/Readers/Biometric Readers. The category information will need to be saved in a separate Category columns in the database.

i. Include an option to collect only “In Stock” and “In Stock Closeout” items. This checkbox is found in the Availability section in the left menu bar.

b. The Update mode will update the pricing and inventory for items already in the Access database. This site uses static links for products. The part numbers are saved in the database. The link is obtained by combining the static link and the part number. For example on part number “BS-SIO2” (which is the first item listed in the category mentioned above) The static link is “http://adiglobal.us/Pages/Product.aspx?pid=BS-SIO2” So all the spider needs to do is combine “http://adiglobal.us/Pages/Product.aspx?pid=” and “BS-SIO2”. Using the ADI\_Example table, in the column PART\_NUM, you can enter any of those part number and go directly to the page.

5. In both modes there are two levels of product details. The first part is in the top of the page. Which is a title (abbreviated description), Model #, ADI #, Price, and Inventory (Price and Inventory are only viewable when logged in). Using the BS-SIO2 as an example again, that date would be

a. VDR\_IT\_DSC – “SECURE I/ O / DR ENCRPT”

b. VDR\_PART (Model #) – “SIO-2”

c. PART\_NUM (ADI #) – “BS-SIO2”

d. Cost (Price) –

i. There is also items marked as “Sale” which I will detail later

e. Inventory

i. I will detail this more later when I give you the login. But for now, you can see in the example database that the spider collects data on the location and the stock and total stock levels.

6. Product Details – In the example database there is a table named ADIChild\_Example. This data comes from the tabs named “Similar Products”, “Overview”, and “Specifications”. There is no standard layout or format used by the site. Some products have a lot of details and others have very little or none. There is also different names for the same data. There are over 600 different field names. Since it is not possible to have that many columns in Access, the best method we could come up with is to create a separate table and create separate records for each field name. This actually seems to work pretty well, but I am open to you suggesting a better method.

7. Time/Date Stamp – In both modes the spider will need to enter the data and time the record was updated in the LastUpdate column.

8. Image Collection – Create an option in both modes to save the image.

a. The current spider save the images in a folder named the same as the manufacturer. This present a problem since some of the manufacturer names have “/” and other non-formatted characters in their name. To resolve the in Web Crawl mode only the first word of the name could be used as the folder name. Or is could replace all non-alphanumeric characters with a “\_”.

i. The Update spider would use a predefined folder name from a specified column in the database. But I prefer the folder be standardized to eliminate duplicate images.

b. The filename the image is saved under is also specified in the database in the Update mode.

c. I welcome your opinion for a better way to do collect images.

9. Program Features

a. Schedules – The ability to preset crawlers to start at a specific time and to have re-occurring jobs that auto start on specified time and day of the week.

b. Multiple Instances – The ability to run multiple browsers windows in order to speed up the crawling speed.

c. Pause – The ability to pause the program from crawling. The program would start where it left off when taken off pause.

d. Ping Rate – The ability to set the max ping rate (pages per minute) that each instance (browser window) will load a page from the site.

e. Auto Detection – The ability for the spider to detect with the page is finished loading, or at least when the required data is loaded on the page. This is used primarily because this sites load time varies from a couple seconds, up to 20 seconds, depending on the time of day. The spider needs to know when it can move forward to the next page.

f. Auto Login – The ability to login into the site automatically.

g. Auto Re-login – The ability to re-login to the site. This is used when it loses connection to the server. Which is also used in conjunction with the Auto Detection feature, which will let it know when it gets logged out (no price and inventory information is displayed)

h. Auto Skip – The ability to skip to the next record when there is an error loading the page. This is used more during the Update mode if a part number is entered that is no longer available. A 404 error, or other type of error is received. The spider will need to move forward to the next record.

i. Adjustable Auto Retry – This is used in conjunction with the Auto Skip feature. If there is a partial page load or other error when loading a page, the spider will try to reload the page for a set number of times before skipping to the next page. For example, 4 attempts.

j. Image Prefix – The ability to add a prefix on the image filename when saving the filename. This is used to know which spider collected the image and which site is came from. We have spiders for several different sites.

k. Email Notification – This is used if there is a major error in the spider. Like if the site is down or the spider cannot login. It will send an email to alert us to the issue.

i. This is not a high priority feature. Let me know if including poses a problem.

l. Ability to select the database and table name for use. This is so we can manually go in through the settings and select the database and name of the table for greater flexibility.

m. Form to enter the website login credentials. – An area in the settings table for entering and changing the login for the site.

n. Task Manager – Used to show pending and completed task. The completed task show the name of the task (the name is entering when setting up the spider) with the date and time the spider completed.

o. Incognito Mode – The ability to spider the site without logging in. Used to collect everything except Pricing and Inventory.

p. I welcome any other features you recommend or have included for past clients.