**Web Spider Project Proposal**

**Phase1 : 15th Sept - 21th Sep : Web Crawler Mode**

1. The UI will list all category. If a category is selected then it will display its sub categories. If a subcategory is selected then sub categories will be displayed. User have option to select any category/subcategory...

Include a By Manufacturer mode also with the ability to select which manufacturer(s) or all.

**OK. Will be done**

2. Two checkboxes: InStock & Stock CloseOut.

There is In Stock items and special order items, although the terminology may change among them. But the In Stock is always known because the Inventory is visible when logged in. Either way, I need the spider in Web Crawler Mode to collect data on all items, not just In Stock. If it runs faster when only collecting In Stock items, then that is fine as an option.

**OK> Crawler mode will crawl all items**

3. A checkbox will be there whether to store image or not. The image will be stored in a folder named as ProductName e.g. B"S-SIO2". This is because product name is unique. There will a table for Product\_Images having columns as (ID, ProductName, ImageName).

What will the ProductName be based on?

**Product Name is the unique code(name) given to each product by website.**

4. A GO button will start parsing and will scrap all items and save into database.

I need a pause button included. I also need the spider to auto detect where it left off, in case of reboot (controlled or not).

**IN 3rd Phase**

5. There will be 10 parallel instances working.

This needs to be controllable, either by selecting the number of instances or the max per minute ping rate. Also include a ping rate schedule so I can throttle up the ping rate at night or on the weekends and lower during the day.

**IN 3rd Phase**

**Phase2 : 22nd Sept - 28th Sep : Update Mode**

1. A login form to store credentials which will used for scrapping inventory information.

Also include a checkbox option in the Web Crawler Mode to be logged in, for price and inventory.

**Will be done. See below point.**

2. A checkox for "Crawler Mode" or "Update Mode".

3. If "Update Mode" is selected and GO is click, It will login through site and will find cost and inventory info for all products in database and update.

Add the option to collect the images in this mode. Actually we normally collect images only in this mode only because it allows us to control the image filename and folder it is saved in. I did not send you the table the current spider uses for the Update mode. There is not much difference except the addition of an Image\_Folder column, which is the name of the folder for all images saved under that manufacturer. And two image name columns with preset image filenames. The current spider adds a filename extension on the end when an image is saved. Which is how we know where there is an image for that item. For example, the image file name in the column will be “123456”. If not image is found for that item, then it stays the same, but when an image is found, the filename changes to “123456.jpg” or 123456.png”, etc. I can use a file rename program to manipulate the image filenames for SEO purpose and then upload to our server with static links, like “Honeywell/123456.png”. It is very important this feature is included.

**OK. I have better option which I explained before. Create a separate table for images. Here columns are (ID, ProductName, ImageName(with extension))**

4. For product additional details, we will use a separate table "ProductAdditionalDetails".

How will this data be stored, similar to now or with more columns for each field type?

**The table structure will be (ID, ProductName, FieldName, FieldValue). So it can supports all fields of a product irrespective of count of parameters.**

5. Auto-Login: If "Forgot Me Not" checkbox is checked then it will save credentials and will start.

The UI will need to auto-detect when it has been logged out or there is a connection problem with the server. Basically by knowing when it is or is not getting the data it requires, such as Price. When not, have it re-attempt the login periodically until successful.

This will need to be saved also in Scheduling so program can auto-start and login when scheduled.

**Whenever we get “Session Expire” page from website, it means that login has expire. If “Forgot Me Not” has been checked then it will login again and start working.**

**Phase3 : 29th Sept - 5th Oct : UI Enhancements**

1. The UI(user interface) will be enhanced to support Scheduling, Multiple Instances Count, pause, Ping Rate, Retry Count, Image Prefix, Email Notification, "Task Manager" will be supported.

**Phase 4: 6th Oct - 12th Oct : Testing**

1. This week is for testing and fixing bugs. This can also be treated a buffer days in case of any technical delays.

**Questions**

1. "Isn’t Incognito Mode" is similar to "Web Crawling Mode".

Web Crawling Mode is used to crawl the entire site to collect new items. The Incognito Mode allows us to scrape it all without worrying about Ping rates, since they only know the IP, which we can change based on where machine is. Then we compare against our existing database, and set the Update spider to update just those records. You can add an option to update just the new items. Otherwise we copy into a separate table and then choose that table to update. The reason I want the option to be logged in during Web Crawling Mode is so when we are not in a hurry to update our database. I can set a low ping rate and let the spider run for a week and gather price and inventory at the same time.

**I am not clear for this point. What is meant by “**The Incognito Mode allows us to scrape it all without worrying about Ping rates, since they only know the IP, which we can change based on where machine is. **”??**

1. "Task Manager" will show list of all tasks and its status.

Have it show Completed and Pending task along with an identifier that shows if the task is reoccurring, in which case it should also show a last completed date. For example, the spider will be set Friday night to run. When we come in Monday we can look at the logged to confirm if it finished. The reoccurring marker also tells us when the spider is set to run again and we can adjust as needed.

**OK.**

3. I have to think about below functionality. Whether this is really needed.

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.

This is used like in question 1 above. So we can set up a small spider scrape. For example, I may crawl the entire site and decide I want to set up a reoccurring spider on a specific set of products. I can copy just those to a table and then set the spider to run on that table. Now, it would be a big help if I can choose different tables and/or databases (or different tables in the same database) depending on the scheduled task. So one task will Web Crawl and save the data in one table, then the Update spider will save its findings in another table.

**OK. Can we modify it like this that only database name should be defined. The table names and table structure will remain same.**

4. Email Notification required SMTP server. Do you have any?

Kind of, we use smtp.gmail.com. We don’t run our own email server. Our server uses Gmail servers for email.

This is not a high priority. If you can included great, but if it is outside of you expertise or would require significant additional cost then it can be omitted. The reason it is on there so when know then there is a failure in the spider. For example we the Friday night task starts, but the ADI website goes down. The spider would email us after a set number attempts to access the site. We would get the email on Saturday and could check on issue and try to fix so the spider can finish by Monday. Otherwise we would not know until Monday and, if it is a 2 day scrape, would not finish until Wednesday.

**OK. This can be done**

**Suggestions**

1. No Auto Detection required. The technology I am using gives us ample time to load page.

The Auto Detection requirement is included because of past problems with our current spider. Whereas due to either poor internet connection or ADI server load, the response time may jump up to 20 seconds or more. And it can vary even within the same minute. One page may load in 3 seconds and the next 15 seconds. The spider would not record all of the data on the long page load and we will not know until we review the data and notice there are several records with no price, inventory, etc. But you are the expert, so if you say it is not needed, then exclude. I just don’t want more issues with the spider and then you have to go back and add in.

**There is a difference between methodology used earlier and Now what I will use. I will parse the page only when its html is complete. So not required.**

2. There will be no display of browser as it is with old spider.

Understood, the results is all I care about. The method is not important to me.

**Other Questions**

1. Can you add to program at a later date the ability to download PDF’s?

**OK> YES can be implemented later.**

1. Can you add to the program at a later date other sites? We currently crawl 6-10 sites periodically. If so, I assume the cost should be less since base program is already created?

**OK> YES can be implemented with other sites. The cost will be only dependent on design and html structure of site.**