

**Team 2**  
**Project Forager**

Project 2  
Sprint 1  
June 26, 2013

by

Alex Veit  
Damon Heard  
Karl Kamdem  
Paul Gimou

A project report submitted for  
SWE 3613 Software Systems Engineering  
Summer 2013

Department of Computer Science and Software Engineering  
Southern Polytechnic State University  
Marietta, Georgia

## TABLE OF CONTENTS

TABLE OF CONTENTS .....	2
1. INTRODUCTION .....	3
1.1Executive Summary .....	3
1.2Project Goals.....	3
1.3Cycle Goals .....	3
2. REQUIREMENTS .....	5
2.1User stories .....	5
3. DESIGN .....	6
3.1System Architecture.....	6
4. RISK ASSESSMENT & MITIGATION.....	7
4.1Obstacles and Risks .....	7
5. CYCLE POSTMORTEM ANALYSIS .....	8
5.1Management Plan Post-Morem Analysis.....	8
5.2Successes .....	8
5.3Failures.....	8
5.4Lessons Learned.....	8
6. TEST PLAN AND PROCEDURES.....	9
7. SYSTEM ADMINISTRATION, INSTALLATION GUIDE & USER MANUAL .....	12
7.1SYSTEM ADMINISTRATION .....	12
7.2INSTALLATION GUIDE .....	13
7.3USER NOTES .....	13
APPENDIX A: SUPPORTING DOCUMENTS .....	14
A.1Correspondence.....	14
A.2Status Reports .....	28
A.3Additional Documentation .....	<b>Error! Bookmark not defined.</b>
APPENDIX B: CYCLE PRESENTATION .....	<b>ERROR! BOOKMARK NOT DEFINED.</b>
APPENDIX C: SOURCE CODE .....	<b>ERROR! BOOKMARK NOT DEFINED.</b>

## 1. INTRODUCTION

### 1.1 EXECUTIVE SUMMARY

Project Forager, almost as its name implies will be a program the scourers the entire domain of SPSU.EDU. Instead of searching for food or provisions the program will look for any errors in links related to images, downloads along with dead or broken links. This information will not only be gathered but the user of the program will also be able to save the results. Saving the results allows one to view it at a later time, compare recent and previous runs, and lastly which will be explained later, rerun results. The Forager will be accessible from anywhere the user can connect to the internet on a computer. This also includes accessing saved results from previous runs. The user will be able to set how deep they want the forager to go, meaning the program does not have to go into every level of the site. This allows the user to specify which areas of the spsu.edu domain to focus on for efficient and quick feedback. With a side by side comparison the user will be able tell which errors were caught on two separate days or runs. The rerun function allows the user to not only go back and rerun previous scans but check the specific errors that came up in the previous scan. This ultimately allows the user to see if those same errors still exist. The Forager with its strong capabilities will be very easy for the user to use.

### 1.2 PROJECT GOALS

The ultimate goal of this project is to check how error proof or vice versa the spsu.edu domain is. To achieve this goal a Web Crawler will need to be built. A Web Crawler is “a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion (Wikipedia.com).” This brings us to the first goal of this project, which entails building a Web Crawler from scratch specifically for the entire spsu.edu website. The second will be outputting the results of this SPSU specific Web Crawler for the user. The Third and last goal of the project is enabling even the average person with little if any technological background to use it. There is more thought and detail that go into these individual goals, but they will instead be discussed in cycle goals.

### 1.3 CYCLE GOALS

During this cycle we intend to get the web crawler portion fully done. This means it will be able to search through all the URL's and images on the site and report errors. Next we will be added functions and features for ease of use. Those contain the ability to save reports for later viewing, filtering results so one can look for a specific page and sorting. So far the program will consist of sorting the error report by number of errors, number of links, error type, alphabetically by page title or URL. More may be added or subtracted to that list. And finally for this cycle we will allow the user to enter a number to decide the breadth of the search. This allows the program to know how many links away from the home page to search until it stops.

## 2. REQUIREMENTS

### 2.1 USER STORIES

- Priority Legend:
  - 1 = High
  - 10 = Low

ID	Name	Story	Priority	Estimated Cost (Hours)
1	Web Account	As a user, I want a unique account on the web from where I can access a web crawler.	1	10
2	List Hyperlinks	I want this crawler - when launched- to be able to scan and retrieve all the links to pages under the spsu.edu domain.	2	20
3	Error Detection	I want the crawler to detect all the html code errors in each page under the spsu.edu domain.	3	25
4	Error Handling	I want the "detected errors" to be stored as reports and displayed in a user friendly manner.	4	10
5	Error Sorting	As a user I would like to sort the "detected errors" by a few different criteria.	7	10
6	Report Comparison	As a user I need to be able to compare previous/ reports with current reports.	8	25
7	App Feedback	As a user I should be able to assess the status of the current scan through some sort of application feedback.	9	10
8	Email Notification	As a user I expect to receive an email notifying me that a scan has ended.	5	5
9	Export Report	As a user I should be able to export any report into a PDF format.	6	10
	<b>TOTAL</b>			<b>125</b>

The team believes that it can turn user stories 1,2, 3 and 4 into working functionality by the end of sprint 1.

### 3. DESIGN

#### 3.1 SYSTEM ARCHITECTURE

IIS Express and .NET framework

The Foragers project settles on any hardware having an internet connection. It uses internet browsers such as Google Chrome, Mozilla or even Internet Explorer.

## 4. RISK ASSESSMENT & MITIGATION

### 4.1 OBSTACLES AND RISKS

Risk assessment

P = Probability

I = Impact

- The use of asp.net and C# is not the correct language for the web crawler and the use of a different coding language may be needed
  - $P = M - I = H$
- Reporting all the broken links to a database will result in issues
  - $P = M - I = L$
- Putting the program over to a server to make it accessible anywhere may result in username malfunction.
  - $P=L - I=M$
- During the runs the program may crash and all current results may be lost.
  - $P = M - I = L$
- During comparison of results, the users screen may not be able to handle the side by side or have it not fit because of screen resolution.
  - $P = L - I = M$
- The results may fail to save to the server.
  - $P=M - I=H$
- Productivity loss to the team due to support requests may be higher than expected.
  - $P = M - I = H$
- New user interface presentation does not satisfy the users leading to rework.
  - $P = L - I = H$
- Achieving an interface compatibility with our product takes longer than planned which increases the effort required for features.
  - $P = M - I = L$

## 5. CYCLE POSTMORTEM ANALYSIS

### 5.1 MANAGEMENT PLAN POST-MOREM ANALYSIS

The team made too many assumptions and ended up allocating very little time per user story. So we ended up devoting a lot of time just figuring out how to recursively crawl the given domain without running out of RAM.

We've only accomplished a fraction of what we set out to do. So it turned out that our size/estimates were once again unrealistic.

### 5.2 SUCCESSES

We were able to successfully setup the local web server using Microsoft's IIS Express & .NET technologies. We developed an algorithm that crawls every single href/link in any given domain, which in extreme cases works best on machines with massive amounts of RAM, but will generally work for most cases.

### 5.3 FAILURES

Not able to setup virtual development environment. Lack of knowledge and perseverance from individuals.

Once again the team failed to work as a unit due to the significant difference between the individuals regarding knowledge of the material at hand. The time management and task delegation was less than bad. A procedure to assess resources, time, and energy necessary for each task is nonexistent.

### 5.4 LESSONS LEARNED

Projects of this magnitude require full commitment of all the team and ample knowledge of the technologies and frameworks at hand. In other words, if an individual's focus is not directed 100% to the project, chances are good that the individual will experience stress, frustration and possibly anger.

Not knowing if each individual can perform at the level stipulated by the requirements is a bit discouraging and is not suited in the SCRUM environment.

A strong SCRUM team is a multifaceted group of forward thinking individuals. This team has lots of learning to do.



## 6. TEST PLAN AND PROCEDURES

Test Cases: Forager

### Test Procedure #1

Step	Action	Expected Response	Notes
1	Start the web crawler	Data shows that web crawler has begun	Web crawler is started in command line for Sprint 1
Web Crawler Began [ ] Pass [ ] Fail			
2	See that a document was made for links	A list of links has populated a document	
3	See that there are links in the document that both pass and have errors	The document has URL's both broken and working	Working URL's are in green broken URL's are in red
List of Broken URL's [ ] Pass [ ] Fail			
Test Procedure #1 [ ] Pass [ ] Fail			

### Test Procedure #2

Step	Action	Expected Response	Notes
1	Start Web Crawler	Data shows that web crawler has begun	Web crawler is started in command line for Sprint 1
2	Cancel the web crawler	Web crawler data stops crawling through the web site	
Web Crawling Stops			

[ ] Pass [ ] Fail			
3	Go to Generated list from canceled web crawl	List was made and able to be seen	
4	List had less links because of the cancelation	List should be small	Don't expect error pages since crawl was canceled
List generation from canceled crawl [ ] Pass [ ] Fail			
Test Procedure #2 [ ] Pass [ ] Fail			

### Test Procedure #3

Step	Action	Expected Response	Notes
1	View Generated List from Test Procedure #1	Test Procedure #1 has been completed and list can be viewed	If Test Procedure 1# has not been completed return and complete it before this test
2	Find a link that has passed in the list	The link will be in green	
3	Copy the Link of the working URL	The working link has been copied	High light the URL and right click to copy
4	Paste the working URL into a web browser and attempt to go to it	The URL is shown and in working order	
Passed Links are Conformed Working [ ] Pass [ ] Fail			
5	Go back to the list and find a link that is in red for not working	Not working link is found in the list	
6	Follow steps 3 and 4 with the non-functioning link	Web page does not work properly	
Failed Links Conformed as Not			

Working [ ] Pass [ ] Fail			
Test Procedure #3 [ ] Pass [ ] Fail			

## 7. SYSTEM ADMINISTRATION, INSTALLATION GUIDE & USER MANUAL

### 7.1 SYSTEM ADMINISTRATION

Functions:

- public partial class \_Default : System.Web.UI.Page()
- private static bool ConstainsHTTP(string url)
- private static string GetDomain(string absolute\_domain)
- private static string Get\_normalized\_Url(string url)
- protected void Page\_Load(object sender, EventArgs e)
- private void Get\_Links(string url)
- private bool Has\_Been\_Visited(object page)
- private void Send\_Email()
- public class DocumentWithLinks
- public DocumentWithLinks(HtmlDocument doc)
- private void GetLinks()
- private void GetReferences()
- private void ParseLink(HtmlNode node, string name)
- public static class CrawlerUtilities
- public static string MapVirtualUrl(string originalURL, string virtURL)
- public static string GetRemotePageContents(string url, out string error)
- public static void SendReport(string url, Dictionary<string, string> badLinks)
- public static string GetAppSetting(string key, string defVal, bool blnRequired, string keyName)
- public static string GetAppSetting(string key, string defVal)
- static public string GetAttributeValue(string rawContent, int eqSignIndex)
- static public string[] GetBetween(string rawContents, string start, string end)
- static public void CleanHtmlComments(ref string html)
- static public void WriteToFile(string fullUrl, string errorMessage)

- private void Initialize()
- private bool OpenConnection()
- private bool CloseConnection()
- public void Insert(string link, string error)
- public void Update()
- public void Delete()
- public int Count()
- public void Backup()
- public void Restore()

## **7.2    INSTALLATION GUIDE**

Launch the Forager website. Log in with an SPSU Email address; Note: if the Email address does not contain @spsu.edu, the log in function won't work. Once logged in, click on the report button to launch a new report about broken links of the spsu.edu domain.

## **7.3    USER NOTES**

Explain in detail any other important information that is relevant to the project and that may be needed by future development teams or system administrators.

## APPENDIX A: SUPPORTING DOCUMENTS

### A.1 CORRESPONDENCE

Karl-----> Group

Hey guys, i took study room 219

Sorry, 221

sup y'all!

I have figured out and implemented how to upload and download files from mysql db. The implementation is in the attached file . The attached file also includes the database file "files.sql" (just a single table for storing the files to support the implementation). You might want to modify the connect.php (connection to the db) so that it matches with the db you have on your local server.

Damon: How far are you. Please look at it and see if u can already integrate that with the UI you have (cuz it is not styled). Have you figured out the email activation stuff yet for login? if not, i could help on that to reduce your workload while you try to do the integration. What d'you think?

Paul & Alex: How far are you guys the db system design that is gonna support our file system? I have started thinking on how we implement the sharing of files between users. If you already have something stable with the db, please let me know so we figure out how to implement sharing.

What do y'all think?

Regards,

Karl.

Ok y'all. We'll make it 3pm at the library or my apartment.

Yeah. My apartment is courtyard 1210. The library is closed

228

218

Damon, after a second thought, d'you think you can start looking at the file system tree implementation? I think with Paul we'll try to come up with a solid database schema and implementation. This might be the basis of

what Alex will work on. So Alex you might want to hold on a lil by working on the UI's glitches. We'll better talk about this tomorrow! What d'yall think?

Hey guys, how are things moving? I just came back and i'm working. On the implementation of the tree structure with the test db!

I'm working on that structure right now! I think i'll be done by tomorrow when we meet! Then we'll integrate it! Just prepare an empty <div> for it

Ok we'll make it 6 then! Is everyone ok with that?

Hey guys, please go online and update your hours on the status report! I wish to send it first thing tomorrow morning

Our schedule depends on Damon

I've integrated the file structure in the system. It works perfectly fine. Now i'm onto sharing!

Hey alex, how far with the server stuff?

Please start without me! I'll be there in 15

mins

Guys, can we make the meeting time 6:30? I'm struggling to have my computer fixed

Yall already there? I was checking on my computer at the store. On my way! Back!

Is it possible to meet a little earlier? Like 11?

Paul we're meeting in your apartment!

I'm in 219

We probably can't have a daily scrum today! Let's all skype at 8. My skype is "karlloickamdem"

Y'all on skype already?

Hey guys, i got the sprint backlog up there! Please add your hours so i can generate the burndown chart for the presentation. Thanks.

Karl <-----> Alex

I will be there in 15 minutes

Ok

I have not gotten home yet I am out for the weekend with my girlfriend I will go back home

on Monday and then I will work on it

Ok. Are you available to meet on monday?

Yes i am

Then i guess we'll try to meet on mon and tues to see where we stand!

Sounds great to me

Ok!

When and where are we meeting tomorrow?

Don't know yet. We can meet at my apartment if the campus is closed. I live on campus

Sounds good, what time you think?

No idea yet!

Well let me know If any one contacts you

Ok!

Damon can meet at about 3

Good with me

Just confirming that we are meeting today at 3 right

Yeah. My apartment is courtyard 1210. The library is closed

I will go to Karl's apartment at 3 o'clock, Damon said he can only make it there at 430

let me know when you send that folder to my email

Done

Thanks

Hey man, can u look at the email activation if u have time. I've tried in vain.

I think it's a feature that should be employment and neck sprint, its pretty convoluted running from a local host

Ok. I get it. I spent all night trying to figure that out. Lol

I got some nice improvements on what we worked on yesterday I'll show it to you today I'll

bring my laptop

Cool! Great! Looking forward to that!

I will be there in 50 minutes



Ok

Ok

Eta?

I am walking into the library right now

Ok ok

Ok

hey guys tonight I'm going to work on better implementing the login functionality so every user can have their own files

Ok, that's cool I'll just make sure that the login is working properly

Ok! Great! We'll get things to work!

where are we meeting guys?

I guess in the library!

anyone already at the library?

I'm omw

I'm in room 221

Ok. Be right there

hey guys, using the cookie has really proven to be useful, I am able to save user information, our software is going to be awesome

Cool I've got everything working nicely as far as the log in and upload/ download goes we just need to implement the file structure

Ok will do

Great!

We will put your file structure in the place of where I am displaying the files right now

Ok!

I'm really having a hard time sending an email activation, how far are you Karl on your file structure?

Just getting started. I've been looking at the code! I'll be setting some test db to pull the info from

I just got here at the school where you guys at?

224

Wsup man?

Have to take care of some personal stuff

Ok!

I can meet today at 6 on campus

hey guys I'm going to be in the library in 3 minutes

221

I updated my hours on the status report I just hope I did it right, this week I didn't put much work into it :(

Ok! I'll check that out! Have you been able to find a free host on the internet to host the project and try the email stuff?

Not yet

Good luck with that man! I feel u. Had a hard time with that too!

what time are we meeting tomorrow guys?

Our schedule depends on Damon

okay guys we meet at school at 11 a.m. tomorrow

Ok!

I'm walking to the library guys

the library's closed guys

I am meeting Karl at his apartment guys

Hey alex could you upload what you have on github!

it is already all up there, except for the database

Ok! Cool

I sent you the db in your spsu email

Ok

Hey, I think there's still room for improvement on this project! I'm really working my ass off to

make this work but i need your help on this. What i need you to do is to fit in Damon's account page (he has sent it by email) with what we already have. Cuz that's what i will need, in order to implement the file structure. D'you think you can get that done by tonight?

No, i can't, i don't have internet where i am

So when can you have it done? Cuz i'm really not good with design!

I can't have it done, can Paul help?

You know the system better than anyone else on the team! Dude we have 4 more days, we can't give up now!

what am I supposed to do with this layout its completely different from our Design plus the code is a mess in completely not organized

Please I need you to fit that left div to the left of the page so i can be able to put the file structure! I can't. If it is confusing, u can always refer to Damon!

I organized that code but I'm not understanding what you want me to do with that <div> what <div> are you talking about?

Ok! This is what i'm saying: i need an empty div to the left of the user home page of the current system so i can place the file structure in it

Do you have the link for the github repository? if you do, all you have to modify is the gooduser.php line 142 to line 157

What about the css? I don't explicitly know how to do that!

as far as the CSS goes for a file structure I didn't write that I took that off the internet, so I also don't know how to tweak it

Ok. I'll figure that out

I will leave that to your discretion

Ok

Perfect, so glad to hear that

Lets use your laptop to present the project

I'm trying to write this piece of code that when a person clicks on the upload button but has not selected any file nothing will happen

Good. I think there are some other user stories we haven't turned into functionality yet: admin account, 2GB/10GB space limit, forgot password etc. these ones will earn us more credit i think. What d'you think?

I agree, the upload button thing is more of a bug fix

Yh! If you could implement (even just the interface) for maybe the admin so i come plug the back end, it'll be a plus for us!

Making an admin profile requires changing the database a little bit

What i think is that the only admin will be entered manually into the db as a normal user. But with the attribute "user\_type" set to "1" (students have 2). From that he'll have his custom privileges: viewing all files in the system and blocking accounts. His GUI will be a little different from the gooduser page too! That's how i see It. Probably no need to modify the db. We might just add an attribute in the users table to be able to block a user's account

Gotcha

Ok

hey guys I think I got a legit online web server for us to host our project if you want

Is it free?

its my professor's web server that I am using for another class

Karam? That'll be cool!

Yes, Karam

I have not deployed it yet, I didn't want to deploy it cuz you said you had some nice implementations

Oh ok! I am done with sharing just now! I'll show it to you guys later today!

okay I am on my way to school I will be there in 20 minutes

Ok!

221

221

On my way!

sorry guys I'm a little late

did you guys already get together?

Not yet. Sincerely, i worked all nite. I'm just waking up

Please start without me. I'll be right there

Lol me too

I'm still at home, no one else has replied

Lol.

Let's make it 10

okay guys we are pushing the meeting to 10

Ok

Ok

guys I'm not sure if I'm going to make it to the school today, you think you can do to meeting without me?

I'm not sure we can do without u! We need everyone on the table. Cuz we need your research and need to leave that place without team assignments ready!

Oh well I'll do my best

Great. Thanks

We can always push it to 7 u know!

Its not the time that matters, but the amount of gas I'm going to spend, I'll have to drive 19 miles to school then another 17 to my girl's house.

Right. I see. Just do your best man! For the sake of the scrum team!

Ok, will do

I'm in 221

I'll be there in a few minutes please

Dude, did u get a chance to setuo the VM?

I'm doing it now

I'm sorry! D'you mind if i do it? Cuz i need a stable platform to begin work on now!, since i have no computer and we meet only monday!

feel free to do it if you want to :) I'll get the vm file with ya later

Cool!

I'm not at home, and i have no internet where I'm at

??

Damon was asking who is going to submit the assignment that is due tonight

It was user stories. We already sent it during our other meeting

Gotcha, cool!

were you able to set up the virtual machine and play around with the .NET framework?

I am heading to school guys I will be there in 40 minutes

I'll be there in 15 minutes

Im in 221

The virtual disk is working but none of the programs are working including Microsoft Visual Studio 2010

Seriously? Damn! We'll look at that in a fee minutes before class!

Ok

Can you please print the status report and bring it to class? I already submitted it only and i'm out of ink! Thanks

Ok

guys I got the HtmlAgilityPack to work!!! now I just need to learn how to use it :)

Great! I got it too! I was able to use it to download a page! Now i have to scan that page!

Great man. Keep up!

I gotta work this Saturday, but i can meet Sunday for this weekend guys

I think coding is the most important. If you have functionality working, good

Gotcha

I'm in 221

How far with the project man?

I got the links using recursion, no multithreading yet

Dude, this will sound awkward but i have a PROPOSITION! I have had such a hard time with c# that i was considering switching to java. Again, this is just a proposition. I found a project that implements a multithreaded web crawler in java! We will have to check for broken links ourselves. You are a better programmer than me! What do you think given the 3 days we have left?

Sounds like a receipt for disaster, lets stick with c#

Ok. You're right!

tell you what... if you want to use java that's cool but you have to have something working by Wednesday I'll try to have something working by Wednesday for c#

We can't work on two different projects.we are a team! I'll continue with c#

Sounds good! :)

Yep!

We are meeting tomorrow at 1pm right guys?

Yh! Is it possible to meet a little earlier?

What time?

Like 11?

what did the guy say about that time?

Sup did you guys establish what time we're meeting?

No one replied to my message yesterday

Ok so the meeting time well be at one

I guess!

hey guys I'm going to be a little late but I'm on my way I'll be there in 20 minutes

can we meet in one of your guys apartments

[www.alexveit.com](http://www.alexveit.com)

My code is getting a stack overflow exception because the recursion is going too deep

I have everything integrated. And it works perfectly with small websites. But with the spsu.edu, there are a bunch of exceptions that abruptly stop the execution

The only one that i am getting here is a stack overflow exception all the other exceptions are being caught and handled

Sure? Anyway u'll have a look at it during the meeting today

hey guys I'm going to be in class in 15 minutes so I will not make it to the daily scrum today, don't forget tonight we have a quiz due online on d2l

Ok! I have successfully crawled 4 websites. I am currently running a scan on spsu

Ok cool!

Hey. The code is on google drive

Cool

I'm on

I couldn't find you on skype Karl

Karlloickamdem!

Give me yours!

I was on with damon earlier

alex.wveit

How far with multithreading?

not looking too good sorry

It's ok

Have u fixed the stack overflow?

No, did that exception pop up with you?

Yeah!

It is blocking me!

Well I'm at work, there is nothing i can do, you can try to figure out a different algorithm to crawl the domain

There's no time for me to start figuring that out now! We present tomorrow! It's better you find a solution to the domain!

Problem!

What do you mean "find a solution to the domain"?



Sorry i meant to say problem! The stack overflow! It's better for us to debug the code now than to find an entire new soln!

Gotcha, i can't do anything now, sorry

Sure. I mean when u get of work and later tonight!

Your a smart individual, you can start thinking on a work around :)

I'm working on data display right now! Can't

Oh well

I'm heading to school I will be there in 30 minutes

I'm in 224

Ok!

Karl <-----> Damon

I been looking at phpadmin and my sql. I can meet with you and work on the integration

Tuesdays

Are you available on monday?

I can meet around 3 for about 30 min

Ok! I'll tell alex! He's ready to meet

Wsup man! We're trying to figure out what to do next. I'd like to know what you have running already so we don't do the same stuff

I'm on my way

Ok man!

Courtyard apartments. Building 1000. Apartment 1210

Yes

Ok!

I won't be able to make it,I have to work over time, i'm going to sneak off to see if skype

Okp

Are u on skype?

We want to meet on campus in 1 hour. Can you?

Yes

Who is going to submit the assignment tonight

What assignment is due tonight?

I think the user stories and the power point, i'm going to double check

We already submitted the user stories, when we had the planning meeting

Oh ok

Dude, can you review your use cases. You had to convert some of our user stories in the document (forager\_user\_stories) into use cases. Can you do it soon ?

Other

I'm taking an exam now,I converted the the user stories I created,I wasn't sure how convert the on

Ok. You're good! We modified them!

Thanks sorry I couldn't modify them, I had to take a bio test right after class

Never mind man. I understand. You're good

Still in class

At what time d'you get out?

720

I'm getting on now

Ok

Damonthegreat3

How far with the login?

I'm done just have to finish the CSS layout

Karl <-----> Paul

We're waiting for u!

Is the virtual disk working properly on your computer?

Gar jai repris ma machine. Jte give la tienne demain if you don't mind!

Ok merci

Im in room 2203 courtyard if u gyuz wannw come. Please Karl dt forget my laptop and the charger

K

Ki a les reponses au quizz?

Not me!

Join us on skype. Now

On es deja go

Tsuip

How far with the project?

Donne moi un truc que je puisse mettre sur le daily scrum minutes!

Gotcha

La reunion sera a 18h! Les gars nont pas voulu do ca le matin!

Ok.

## **A.2 STATUS REPORTS**

# SWE 3613 Status Report

Cover Sheet  
Summer 2013

Group Name	Group 2	Report Date	6/25/2013
Project Name	Forager	Team Members	Alex Veit, Damon Heard, Paul Antoine Gimou, Karl-Loic Kamdem
Project #	2	Sprint #	Sprint 1
Executive Summary	<p>Forager is a web application that crawls a given domain and retrieves all broken html tags so they can be corrected in the future. Project Forager, almost as its name implies will be a program that scours the entire domain of SPSU.EDU. Instead of searching for food or provisions the program will look for any errors in links related to images, downloads along with dead or broken links. This information will not only be gathered but the user of the program will also be able to save the results. Saving the results allows one to view it at a later time, compare recent and previous runs, and lastly which will be explained later, rerun results. The Forager will be accessible from anywhere the user can connect to the internet on a computer. This also includes accessing saved results from previous runs.</p>		
Cycle Intent & Goals	<p>Fix stack overflow exception due to recursion. Ability to crawl recursively any given domain and to get all broken links, images, files, js and css files under that domain. At the end of this process, an email notification should be sent to the user to confirm the successful crawl. Store information about errors in a database and display it to the user.</p>		

SWE 3613 Status Report (Page 2)							
Requirements							

Group Name:	Group 2	Project Name:	Forager
Members:	Alex Damon Karl Paul	Report Date:	6/25/2013

ID#	User Story, Use Case, or Requirement	Planned			Actual		
		Cycle planned for completion	Total planned hours	Planned hours this cycle	Status	Actual hours this cycle	Total hours
1	As a user, I want a unique account on the web from where I can access the web crawler.	1	10	10	Completed	11	11
2	I want this crawler - when launched- to be able to scan and retrieve all the links to pages under the spsu.edu domain.	1	20	20	Completed	61.5	61.5
3	I want the crawler to detect all the html code errors in each page under the spsu.edu domain.	1	25	25	Completed	28	28
4	I want the "detected errors" to be stored as reports and displayed in a user friendly manner.	1	10	10	Completed	3	3
5	As a user I would like to sort the "detected errors" by a few different criteria.	2	10	10	Unstarted		
6	As a user I need to be able to compare previous/ reports with current reports.	2	25	25	Unstarted		
7	As a user I should be able to assess the status of the current scan through some sort of application feedback.	2	10	10	Unstarted		
8	As a user I expect to receive an email notifying me that a scan has ended.	1	5	5	Completed	2	2
9	As a user I should be able to export any report into a PDF format.	2	10	10	unstarted		
		Planned Total	125	125	Actual Total	105.5	105.5

**IMPORTANT: File naming instructions**

Name this file in the following manner: YYYYMMDD\_TEAM\_NAME\_HERE

Example: 20120911\_Group1.pdf

List Data - De

Completed 1

Discarded 2

Development 3

Testing 4

Design 5

Unstarted 6







## Contributions

<b>Instructions:</b>	List all source code files in the project, indicate the total number of lines of code(LOC) and the number of logical lines of code (LLOC, which excludes comments, blanks lines, or any line that does not contain an executable statement), and then indicate which team member(s) contributed to the code in the file.
----------------------	--

File or Artifact	Total LOC	LLOC	Member 1	Member 2	Member 3	Member 4
			Alex	Damon	Karl	Paul
Product Backlog	50	40	10	10	10	10
Default.aspx.cs	450	400	300	0	30	0
Database.cs	130	100	0	0	100	0
CrawlerUtilities.cs	350	300	0	0	180	0
Totals:	980	840	310	10	320	10

# SWE 3613 Status Report (Page 5)

## Timesheet

**Group Name:** Group 2 **Project Name:** Forager  
**Members:** Alex Damon Karl Paul **Report Date:** 6/25/2013

**Instructions** Indicate on which days and for how many hours each team member worked for the past week. Next, indicate the contributions of each team member for the past week.

### Hours Worked

	Wednesday	Thursday	Friday	Saturday	Sunday	Monday	Tuesday	Total
	6/19/2013	6/20/2013	6/21/2013	6/22/2013	6/23/2013	6/24/2013	6/25/2013	
Alex	0	5	3	0	7	3	3	21.0
Damon	0	3	3	8	3	4	8	29.0
Karl	5	6	0	9	9	2	5	36.0
Paul	0	2	1	2	0	0	3	8.0
Total	5.0	16.0	7.0	19.0	19.0	9.0	19.0	94.0

### Major Contributions / Tasks Performed

**Alex** Implemented the recursive algorithms to crawl a given domain

**Damon** I created the login page using LoginView Control. The user doesnt choose which view is used, Instead,the view is set based on the authentication status of the user.

**Karl** Build algorithm to validate all html links and references. Build temporary database schema to store results from the domain crawling. Write code to send email notification to user when crawling and validation jobs end. Test application on spsu.edu domain and other smaller domains for quick results. Integration and integration testing

**Paul**

## APPENDIX A: CYCLE PRESENTATION

# Forager - Sprint 1

## Group 2

Paul Antoine Gimou

Alexander Veit

Karl-Loic Kamdem

Damon Heard

# Introduction

The logo for Southern Polytechnic State University (SPSU) is a green hexagon with the letters "SPSU" in white.

- Forager is a web application that crawls a given domain and retrieves all broken html tags so they can be corrected in the future
- Saving the results allows one to view it at a later time, compare recent and previous runs
- The Forager will be accessible from anywhere the user can connect to the internet on a computer

## **Sprint Planning Meeting - Part 1**

- During the first part of the meeting, the Customer presented the team with the product vision and a product backlog
- From the product backlog, the team chose to complete items 1, 2, 3, 4 and 8.

# Product Backlog

SPSU

ID	Name	Story	Priority	Estimated Cost (Hours)
1	Web Account	As a user, I want a unique account on the web from where I can access a web crawler.	1	10
2	List Hyperlinks	I want this crawler - when launched- to be able to scan and retrieve all the links to pages under the spsu.edu domain.	2	20
3	Error Detection	I want the crawler to detect all the html code errors in each page under the spsu.edu domain.	3	25
4	Error Handling	I want the "detected errors" to be stored as reports and displayed in a user friendly manner.	4	10
5	Error Sorting	As a user I would like to sort the "detected errors" by a few different criteria.	7	10
6	Report Comparison	As a user I need to be able to compare previous/ reports with current reports.	8	25
7	App Feedback	As a user I should be able to assess the status of the current scan through some sort of application feedback.	9	10
8	Email Notification	As a user I expect to receive an email notifying me that a scan has ended.	5	5
9	Export Report	As a user I should be able to export any report into a PDF format.	6	10
	<b>TOTAL</b>			<b>125</b>

## **Sprint Planning Meeting - Part 2**

- The team met to figure out how it will turn the selected product backlog into functionality
- The output of this meeting was the sprint 1 backlog



# SCRUM FLOW - Sprint 1

SPSU

Sprint 2 Backlog		Hours of work remaining		
Task Description	Responsible	Status(Not started/In progress/Completed)	Day 1	Day 2
Setup Virtual Machine for common development environment	Karl	Completed	2	
Download, install and interconnect all development tools (IDE, Database and server)	Karl	Completed	3	
Create a user login page with unique account	Damon	Completed		
Develop crawler algorithm using recursion	Alex	Development	25	
Optimize crawl algorithm with multi threading	Alex	Development	20	
Build algorithm to validate all html links and refences	Karl	Completed	20	
Build temporary database schema to store results from the domain crawling	Karl	Completed	1	
Retrieve data from source and display to the user	Karl	Completed	2	
Develop master template for website	Paul	In progress	18	
Write code to send email notification to user when crawling and validation jobs end	Karl/Alex	Completed	2	
Test application on spsu.edu domain and other smaller domains for quick results	Karl	Completed (Outcome is positive)	5	
Integration and integration testing	Karl	Completed	3	
		TOTAL		

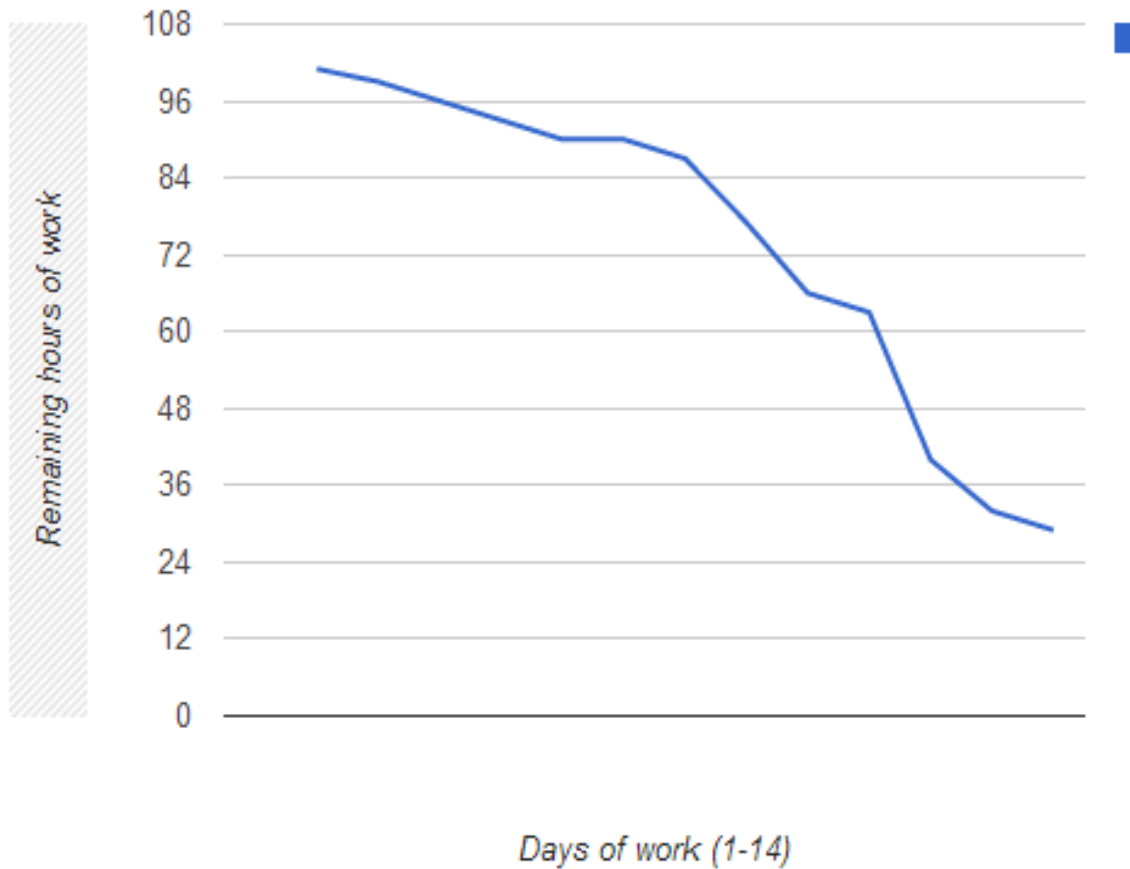
## Daily Scrum Meetings

- A total of four Daily Scrum meetings were held during this cycle to help synchronize its progress.
- During these meetings, each team member - one after the other - answered three questions:
  - What have you done since the last Daily Scrum?
  - What will you do before the next Daily Scrum?
  - What impedes you from performing your work as effectively as possible?

# Actual Sprint 1

SPSU

**Burndown chart**



# PostMortem & Performance Analysis

SPSU

- The team made too many assumptions and ended up allocating very little time per user story
- The actual hours of work ended up being a lot more than the team had planned
- Estimated Hours of work: 60
- Actual Hours of work: 100
- Tasks were not divided into the smallest possible units. This resulted in poor cost estimations

# PostMortem & Performance Analysis

The logo for Southern Polytechnic State University (SPSU) is a green hexagon with the letters "SPSU" in white.

## Successes

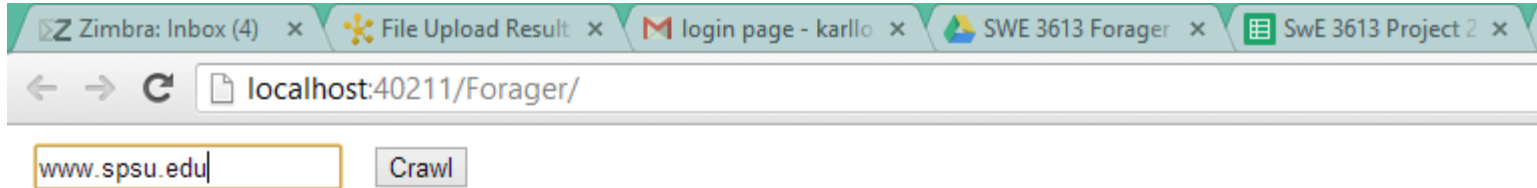
- We developed an algorithm that crawls every single href/link and returns detected errors in any given domain, although it is not time and cost effective

## Failures

- A procedure to assess resources, time, and energy necessary for each task is nonexistent.
- Not able to setup virtual development environment
- The team failed to work as a unit due to the significant difference between the individuals regarding knowledge of the material at hand.

# Demo Screenshots

SPSU



SOUTHERN  
POLYTECHNIC  
STATE UNIVERSITY





links.txt | links.txt | errors.txt | brokenlinks.txt

41987 <http://www.spsu.edu/?view=day&date=20130617>

41988 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10566-10565&pstat=AC&instStartTime=1371441600>

41989 <http://www.spsu.edu/?view=day&date=20130618>

41990 <http://www.spsu.edu/?view=day&date=20130619>

41991 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10568-10567&pstat=AC&instStartTime=1371614400>

41992 <http://www.spsu.edu/?view=day&date=20130620>

41993 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a9057-9056&pstat=AC&instStartTime=13717008000>

41994 <http://www.spsu.edu/?view=day&date=20130621>

41995 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10673-10672&pstat=AC&instStartTime=1371787200>

41996 <http://www.spsu.edu/?view=day&date=20130622>

41997 <http://www.spsu.edu/?view=day&date=20130623>

41998 <http://www.spsu.edu/?view=day&date=20130624>

41999 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10675-10674&pstat=AC&instStartTime=1372046400>

42000 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10611-10610&pstat=AC&instStartTime=1372046400>

42001 <http://www.spsu.edu/?view=day&date=20130626>

42002 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10570-10569&pstat=AC&instStartTime=1372219200>

42003 <http://www.spsu.edu/?view=day&date=20130627>

42004 <http://www.spsu.edu/?view=day&date=20130628>

42005 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10695-10694&pstat=AC&instStartTime=1372392000>

42006 <http://www.spsu.edu/?view=day&date=20130629>

42007 <http://www.spsu.edu/?view=day&date=20130630>

42008 <http://www.spsu.edu/?view=day&date=20130701>

42009 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10614-10613&pstat=AC&instStartTime=1372651200>

42010 <http://www.spsu.edu/?view=day&date=20130702>

42011 <http://www.spsu.edu/?view=day&date=20130703>

42012 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a10733-10732&pstat=AC&instStartTime=1372824000>

42013 <http://www.spsu.edu/?view=day&date=20130704>

42014 <http://www.spsu.edu/?view=month&action=view&invId=2b3749ef-db92-4d1b-9160-024fa9776302%3a9071-9070&pstat=AC&instStartTime=13729104000>

42015 <http://www.spsu.edu/?view=day&date=20130705>

42016 <http://www.spsu.edu/?view=day&date=20130706>

42017 <http://www.spsu.edu/home/calendar@spsu.edu/Registrar> Calendar.ics

42018



```

945 http://www.spsu.edu/help/coursereconsearch.htm "The remote server returned an error: (404) Not Found."
946 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=bbSearch "The remote server returned an error: (404) Not Found."
947 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=sbSearch "The remote server returned an error: (404) Not Found."
948 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=rbSearch "The remote server returned an error: (404) Not Found."
949 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=pbLogon "The remote server returned an error: (404) Not Found."
950 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=REQNOBIB "The remote server returned an error: (404) Not Found."
951 http://www.spsu.edu/cgi-bin/newbooks.cgi "The remote server returned an error: (404) Not Found."
952 http://www.spsu.edu/remote.htm "The remote server returned an error: (404) Not Found."
953 http://www.spsu.edu/help/contents.htm "The remote server returned an error: (404) Not Found."
954 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?PAGE=bbSearch&SEQ=20130625205705&PID=4DwAFY9LYm6uk2ZvevYCV-NbTMqP "The remote server returned a
955 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?PAGE=sbSearch&SEQ=20130625205705&PID=4DwAFY9LYm6uk2ZvevYCV-NbTMqP "The remote server returned a
956 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?sel=Preferences&PAGE=LOGON&SEQ=20130625205705&PID=4DwAFY9LYm6uk2ZvevYCV-NbTMqP "The remote s
957 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?PAGE=LOGON&SEQ=20130625205705&PID=4DwAFY9LYm6uk2ZvevYCV-NbTMqP "The remote server returned a
958 http://www.spsu.edu/vufind/Search/History "The remote server returned an error: (404) Not Found."
959 http://www.spsu.edu/vufind/Help/Home?topic=usqsearch "The remote server returned an error: (404) Not Found."
960 http://www.spsu.edu/vufind/MyResearch/Home "The remote server returned an error: (404) Not Found."
961 http://www.spsu.edu/javascript:self.print(); "The remote server returned an error: (404) Not Found."
962 http://www.spsu.edu/library/Departments/ILL/ILL.html "The remote server returned an error: (404) Not Found."
963 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=Exit "The remote server returned an error: (404) Not Found."
964 http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First "The remote server returned an error: (404) Not Found."
965 http://www.spsu.edu/./userlog/detail.asp "The remote server returned an error: (404) Not Found."
966 "http://www.spsu.edu/javascript:var w=window.open('help/RefWorks2.htm#Welcome.htm', 'RWHelp');if(w)w.focus();if(w)w.location.reload();if(
967 http://www.spsu.edu/javascript:void(0); "The remote server returned an error: (404) Not Found."
968 http://www.spsu.edu/./RWAthens "The remote server returned an error: (404) Not Found."
969 https://www.refworks.com/RWShibboleth?providerId=https://shibboleth.edgehill.ac.uk/shibboleth "The remote server returned an error: (50
970 https://www.refworks.com/RWShibboleth?providerId=https://shib.fortlewis.edu/idp/shibboleth "The remote server returned an error: (500) I
971 https://www.refworks.com/RWShibboleth?providerId=https://shib2idp.rgu.ac.uk/idp/shibboleth "Too many automatic redirections were attempt
972 https://www.refworks.com/RWShibboleth?providerId=https://www.rediris.es/sir/uniriojaidd "The remote server returned an error: (401) Unaut
973 https://www.refworks.com/RWShibboleth?providerId=https://idp.bham.ac.uk/shibboleth "The remote name could not be resolved: 'shibbolethid
974 https://www.refworks.com/RWShibboleth?providerId=https://shibidp.uclan.ac.uk/idp/shibboleth "The remote server returned an error: (500) I
975 https://www.refworks.com/RWShibboleth?providerId=https://dmz-shib-dg-01.dmz.roehampton.ac.uk/idp/shibboleth "The remote server returned a
976 974 Failed Links

```



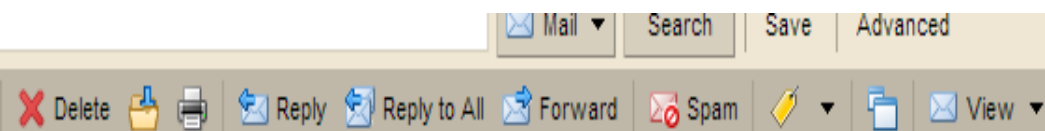
Query 1 x

Filter: File: Autosize:

link	error
<a href="http://www.spsu.edu/browse.php">http://www.spsu.edu/browse.php</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/ask.spsu.edu/login.ph...">http://www.spsu.edu/ask.spsu.edu/login.ph...</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/ask.spsu.edu/index.php">http://www.spsu.edu/ask.spsu.edu/index.php</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/browse.php?tid=10775">http://www.spsu.edu/browse.php?tid=10775</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/browse.php?tid=10804">http://www.spsu.edu/browse.php?tid=10804</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/vufind/Search/Home">http://www.spsu.edu/vufind/Search/Home</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/vufind/Search/Advanced">http://www.spsu.edu/vufind/Search/Advanced</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/vufind/Browse/Home">http://www.spsu.edu/vufind/Browse/Home</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/webvoy.htm">http://www.spsu.edu/webvoy.htm</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?...">http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?...</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?...">http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?...</a>	The remote server returned an error: (404) Not Found.
<a href="http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?...">http://www.spsu.edu/cgi-bin/Pwebrecon.cgi?...</a>	The remote server returned an error: (404) Not Found.

report 9


Read Only



f 2135 mes

## Forager Report

June 25, 2013 4:05 PM

▼ From:  group2forager@gmail.com

To: kkamdem@spsu.edu

The page <http://www.groupeheci.ac.ma> contains 2 invalid links:

'<http://www.groupeheci.ac.ma/index.php/notre-reseau?id=5:option-marketing&catid=16:master>' is invalid, error was:

The remote server returned an error: (404) Not Found.

'<http://www.groupeheci.ac.ma/index.php/notre-reseau/heci-reseau-maroc?id=6:option-ingenierie-et-management-de-projet&catid=16:master>' is invalid, error was: The remote server returned an error: (404) Not Found.

## APPENDIX C: SOURCE CODE

### Default.aspx.cs

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Web;
using System.Web.UI;
using System.Web.UI.WebControls;
using System.Data.Odbc;
using System.Configuration;
using HtmlAgilityPack;
using System.IO;
using System.Text;
using System.Net;
using TestCrawler;

using System.Collections;
using System.Net.Mail;
using System.Net.Security;
using System.Security.Cryptography.X509Certificates;

public partial class _Default : System.Web.UI.Page
{
    /*DO NOT EVEN THINK OF DELETING Karl's old code. DO NOT DELETE!!!
    protected void Page_Load(object sender, EventArgs e)
    {
        try
        {
            using (OdbcConnection connection = new
OdbcConnection(ConfigurationManager.ConnectionStrings["Forager_Connection"].ConnectionString))
            {
                connection.Open();
                using (OdbcCommand command = new OdbcCommand("select * from team", connection))

                using (OdbcDataReader reader = command.ExecuteReader())
                {
                    while (reader.Read())
                    {
                        Response.Write(reader["name"].ToString() + "<br />");
                    }
                    reader.Close();
                }
                connection.Close();
            }
        }
        catch (Exception ex)
        {
            Response.Write("Error: " + ex.Message);
        }
    }
}
```

```

}
*/

ArrayList _references = new ArrayList();
ArrayList _second_references = new ArrayList();
string _absolute_domain;
string _domain;

//string _path = Directory.GetCurrentDirectory() + "\\result.txt";
string _path = @"C:\Users\karlloic\Documents\links.txt";

private static bool ConstainsHTTP(string url)
{
    return (url.Contains("http://") || url.Contains("https://"));
}

private static string GetDomain(string absolute_domain)
{
    string d;
    if (absolute_domain.Contains("http://"))
        d = absolute_domain.Substring(7);
    else
        d = absolute_domain;

    int index = d.Length - 1;

    for (int i = 0; i < d.Length; i++)
    {
        if (d[i] == '/')
        {
            index = i - 1;
        }
    }

    string ret = null;
    int dot_count = 0;
    int len = 0;
    for (int i = index; i >= 0; i--)
    {
        if (d[i] == '.')
        {
            dot_count++;
            if (dot_count == 2)
            {
                ret = d.Substring(i + 1, len);
                break;
            }
        }
    }
}

```

```

        len++;
    }

    return ret;
}

private static string Get_normalized_Url(string url)
{
    string ret;

    if (url.Length >= 7 && ConstainsHTTP(url))
        ret = url;
    else
        ret = url.Insert(0, "http://");

    return ret;
}

protected void Page_Load(object sender, EventArgs e)
{
    using (StreamWriter sw = File.CreateText(_path))
    {
        sw.WriteLine("Links:");
    }
}

//when the user clicks the button on the website this gets called
protected void Button1_Click1(object sender, EventArgs e)
{
    _absolute_domain = Get_normalized_Url(TextBox1.Text);
    if (_absolute_domain != null)
    {
        _domain = GetDomain(_absolute_domain);
        if (_domain != null)
        {
            Get_Links(_absolute_domain);

            //not using for now
            Write_to_File();

            //Send_Email();

            /*

            *
            *
            *

```

```

//Call Karls code sending all the spsu.edu links (href) as parameter
karls_stuff(_references); <<<<===== Karl call your code here
*/
//done. now check each link:
string strErrorMessage;
Dictionary<string, string> arrFailedLinks = new Dictionary<string, string>();

string tmp;
foreach (string line in _references)
{
    if (line.Length > 0)
    {
        if (ConstainsHTTP(line))
            tmp = line;
        else
        {
            if (line.First().Equals('/'))
                tmp = line.Insert(0, _absolute_domain);
            else
                tmp = line.Insert(0, _absolute_domain + "/");
        }
        _second_references.Add(tmp);
    }
}

foreach (string strHref in _second_references)
{
    string strFullUrl = CrawlerUtilities.MapVirtualUrl(_absolute_domain, strHref);
    string dummyContents = CrawlerUtilities.GetRemotePageContents(strFullUrl, out
strErrorMessage);
    if (strErrorMessage.Length > 0)
    {
        Console.WriteLine(string.Format("URL '{0}' is invalid or server is down
({1})", strFullUrl, strErrorMessage));
        arrFailedLinks.Add(strFullUrl, strErrorMessage);
        CrawlerUtilities.WriteToFile(strFullUrl, strErrorMessage);
    }
    else
    {
        Console.WriteLine("URL '" + strFullUrl + "' is valid!");
    }

    //CrawlerUtilities.SendReport(url, arrFailedLinks);
}

if (arrFailedLinks.Count > 0)
{

```

```

        Console.WriteLine("Found " + arrFailedLinks.Count + " invalid links, trying to
send report.");
        try
        {
            CrawlerUtilities.SendReport(_absolute_domain, arrFailedLinks);
            CrawlerUtilities.WriteToFile(arrFailedLinks.Count.ToString(), " Failed
links");
        }
        catch (Exception ex)
        {
            Console.WriteLine("Failed to send report: " + ex.ToString());
            CrawlerUtilities.WriteToFile("Failed to send report: ", ex.ToString());
        }
    }
    else
    {
        try
        {
            CrawlerUtilities.SendReport(_absolute_domain, arrFailedLinks);
            CrawlerUtilities.WriteToFile(arrFailedLinks.Count.ToString(), " Failed
links");
        }
        catch (Exception ex)
        {
            Console.WriteLine("Failed to send report: " + ex.ToString());
            CrawlerUtilities.WriteToFile("Failed to send report: ", ex.ToString());
        }
    }

    /*
    *
    *
    *
    *
    */

    string msg = "alert('Successful Forage');";
    Page.ClientScript.RegisterStartupScript(GetType(), "msgbox", msg, true);
}
else
{
    string msg = "alert('Oops... Something went wrong with the input');";
    Page.ClientScript.RegisterStartupScript(GetType(), "msgbox", msg, true);
}
}
else
{
    string msg = "alert('Oops... Something went wrong with the input');";

```



```
Page.ClientScript.RegisterStartupScript(GetType(), "msgbox", msg, true);
```

```
}
```

```
}
```

```
private void Get_Links(string url)
```

```
{
```

```
    HtmlWeb hw = new HtmlWeb();
```

```
    HtmlDocument doc = hw.Load(url);
```

```
    DocumentWithLinks nwl = new DocumentWithLinks(doc);
```

```
    for (int i = 0; i < nwl.References.Count; i++)
```

```
    {
```

```
        if (!Has_Been_Visited(nwl.References[i]))
```

```
        {
```

```
            //this statement does not append complete url to _references- just the remining
```

```
            _references.Add(nwl.References[i]);
```

```
            //this is used just to print complete links to the file (_path) as they are got  
            using (StreamWriter sw = File.AppendText(_path))
```

```
            {
```

```
                string tmp;
```

```
                if (nwl.References[i].ToString().Length > 0)
```

```
                {
```

```
                    if (ConstainsHTTP(nwl.References[i].ToString()))
```

```
                        tmp = nwl.References[i].ToString();
```

```
                    else
```

```
                    {
```

```
                        if (nwl.References[i].ToString().First().Equals('/'))
```

```
                            tmp = nwl.References[i].ToString().Insert(0, _absolute_domain);
```

```
                        else
```

```
                            tmp = nwl.References[i].ToString().Insert(0, _absolute_domain +
```

```
"/");
```

```
                    }
```

```
                    sw.WriteLine(tmp);
```

```
                }
```

```
            }
```

```
string temp;
```

```
if (!ConstainsHTTP(nwl.References[i].ToString()))
```

```
{
```

```
    temp = _absolute_domain;
```

```
    if (nwl.References[i].ToString()[0] == '/')
```

```
        temp += nwl.References[i].ToString();
```

```
    else
```

```
        temp += "/" + nwl.References[i].ToString();
```

```

    }
    else
        temp = nwl.References[i].ToString();

    if (temp.Contains(_domain))
    {
        try
        {
            Get_Links(temp);
        }
        catch (Exception ex)
        {
            string msg = "alert('Oops... ' + ex.Message + '');";
            Page.ClientScript.RegisterStartupScript(GetType(), "msgbox", msg, true);
        }
    }
}

}

private bool Has_Been_Visited(object page)
{
    for (int i = 0; i < _references.Count; i++)
    {
        if (_references[i].Equals(page))
            return true;
    }
    return false;
}

private void Send_Email()
{
    SmtpClient client = new SmtpClient();
    client.Port = 587;
    client.Host = "smtp.gmail.com";
    client.EnableSsl = true;
    client.Timeout = 10000;
    client.DeliveryMethod = SmtpDeliveryMethod.Network;
    client.UseDefaultCredentials = false;
    client.Credentials = new System.Net.NetworkCredential("group2forager@gmail.com",
"honeycomb");

    MailMessage mm = new MailMessage("donotreply@domain.com", "aveit@spsu.edu", "Forager
Report", "test body");
    mm.BodyEncoding = UTF8Encoding.UTF8;
    mm.DeliveryNotificationOptions = DeliveryNotificationOptions.OnFailure;

```

```

//add following code before smtpClient.Send()

client.EnableSsl = true;

ServicePointManager.ServerCertificateValidationCallback = delegate(object s,
X509Certificate certificate, X509Chain chain, SslPolicyErrors sslPolicyErrors) { return true; };
client.Send(mm);
}

private void Write_to_File()
{
    string name = @"C:\Users\karlloic\Documents\links.txt";
    using (System.IO.StreamWriter file = new System.IO.StreamWriter(name))
    {
        string tmp;
        foreach (string line in _references)
        {
            if (line.Length > 0)
            {
                if (ConstainsHTTP(line))
                    tmp = line;
                else
                {
                    if (line.First().Equals('/'))
                        tmp = line.Insert(0, _absolute_domain);
                    else
                        tmp = line.Insert(0, _absolute_domain + "/");
                }

                file.WriteLine(tmp);
            }
        }
    }
}

/// <summary>
/// Represents a document that needs linked files to be rendered, such as images or css files,
and points to other HTML documents.
/// </summary>
public class DocumentWithLinks
{
    private ArrayList _links;
    private ArrayList _references;
    private HtmlDocument _doc;

    /// <summary>
    /// Creates an instance of a DocumentWithLinkedFiles.
    /// </summary>

```

```
/// <param name="doc">The input HTML document. May not be null.</param>
```

```
public DocumentWithLinks(HtmlDocument doc)
```

```
{
    if (doc == null)
    {
        throw new ArgumentNullException("doc");
    }
    _doc = doc;
    GetLinks();
    GetReferences();
}
```

```
private void GetLinks()
```

```
{
    _links = new ArrayList();
    HtmlNodeCollection atts = _doc.DocumentNode.SelectNodes("//*[@background or @lowsrc or @src or @href]");
    if (atts == null)
        return;

    foreach (HtmlNode n in atts)
    {
        ParseLink(n, "background");
        ParseLink(n, "href");
        ParseLink(n, "src");
        ParseLink(n, "lowsrc");
    }
}
```

```
private void GetReferences()
```

```
{
    _references = new ArrayList();
    HtmlNodeCollection hrefs = _doc.DocumentNode.SelectNodes("//a[@href]");
    if (hrefs == null)
        return;

    foreach (HtmlNode href in hrefs)
    {
        if (!_references.Contains(href.Attributes["href"].Value))
            _references.Add(href.Attributes["href"].Value);
    }
}
```

```
private void ParseLink(HtmlNode node, string name)
```

```
{
    HtmlAttribute att = node.Attributes[name];
    if (att == null)
        return;
}
```

```

    // if name = href, we are only interested by <link> tags
    if ((name == "href") && (node.Name != "link"))
        return;
    if (!_links.Contains(att.Value))
        _links.Add(att.Value);
}

```

```

/// <summary>
/// Gets a list of links as they are declared in the HTML document.
/// </summary>

```

```

public ArrayList Links
{
    get
    {
        return _links;
    }
}

```

```

/// <summary>
/// Gets a list of reference links to other HTML documents, as they are declared in the
HTML document.

```

```

/// </summary>
public ArrayList References
{
    get
    {
        return _references;
    }
}

```

```

}

```

```

}

```

### CrawlerUtilities.cs

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Net.Mail;
using System.Configuration;
using System.Net;
using MySql.Data.MySqlClient;

using System.Net.Security;
using System.Security.Cryptography.X509Certificates;

namespace TestCrawler
{
    public static class CrawlerUtilities
    {
        public static string MapVirtualUrl(string originalURL, string virtURL)
        {
            //maybe not virtual?
            if (virtURL.StartsWith("http://") || virtURL.StartsWith("https://"))
                return virtURL;

            string strCleanOriginal = originalURL.ToLower();
            string prefix = strCleanOriginal.StartsWith("http://") ? ("http://") :
(strCleanOriginal.StartsWith("https://") ? "https://" : "");
            int index = strCleanOriginal.IndexOf("?");
            if (index > 0)
                strCleanOriginal = strCleanOriginal.Substring(0, index);
            string[] parts = strCleanOriginal.Replace("http://", "").Replace("https://",
"").Split('/');
            string root = parts[0];
            string retVal = prefix + root;
            if (virtURL.StartsWith("?"))
                virtURL = parts[parts.Length - 1] + virtURL;
            if (!virtURL.StartsWith("/"))
            {
                for (int i = 1; i < parts.Length - 1; i++)
                {
                    retVal += "/" + parts[i] + "/";
                }
            }
            retVal += virtURL;

            return retVal;
        }

        public static string GetRemotePageContents(string url, out string error)
```

```

{
    error = string.Empty;
    string contents = string.Empty;
    using (WebClient client = new WebClient())
    {
        client.Headers[HttpRequestHeader.UserAgent] = "Mozilla/4.0 (compatible; MSIE 7.0;
Windows NT 5.1; Trident/4.0)";
        try
        {
            contents = client.DownloadString(url);
        }
        catch (Exception ex)
        {
            error = ex.Message;
        }
    }
    return contents;
}

public static void SendReport(string url, Dictionary<string, string> badLinks)
{
    /*
    string receipientEmail = GetAppSetting("ReceipientEmail", "", true, "ReceipientEmail");
    string senderName = GetAppSetting("SenderName", "");
    string senderEmail = GetAppSetting("SenderEmail", "", true, "SenderEmail");
    string smtpServerAddress = GetAppSetting("SmtpServerAddress", "localhost");
    string subject = GetAppSetting("ReportEmailSubject", "Found some bad links");
    string smtpPort = GetAppSetting("SmtpPort", "");

    //format?
    if (subject.IndexOf("{0}") > 0)
        subject = subject.Replace("{0}", url);

    string fullSender = (senderName.Length > 0) ? string.Format("{0}<{1}>", senderName,
senderEmail) : senderEmail;
    SmtpClient client = new SmtpClient(smtpServerAddress);
    client.DeliveryMethod = SmtpDeliveryMethod.PickupDirectoryFromIis;
    if (smtpPort.Length > 0)
        client.Port = Int32.Parse(smtpPort);

    StringBuilder body = new StringBuilder();
    body.Append("The page ").Append(url).Append(" contains
").Append(badLinks.Count).Append(" invalid links:").Append(Environment.NewLine);
    foreach (string badUrl in badLinks.Keys)
    {
        body.AppendFormat("' {0}' is invalid, error was: {1}{2}", badUrl, badLinks[badUrl],
Environment.NewLine);
    }
}

```

```

        using (MailMessage message = new MailMessage(fullSender, receipientEmail, subject,
body.ToString()))
        {
            try
            {
                client.Send(message);
                Console.WriteLine("Email sent successfully to " + receipientEmail);
            }
            catch (Exception ex)
            {
                Console.WriteLine("Failed to send email: " + ex.Message);
            }
        }
    }
    */

    /*
var fromAddress = new MailAddress("karlloic@gmail.com", "Group 2 Forager");
var toAddress = new MailAddress("kkamdem@spsu.edu", "Karl-Loic");
const string fromPassword = "christelle11";
const string subject = "Forager Report";
//const string body = "Body";

StringBuilder body = new StringBuilder();
body.Append("The page ").Append(url).Append(" contains
").Append(badLinks.Count).Append(" invalid links:").Append(Environment.NewLine);
foreach (string badUrl in badLinks.Keys)
{
    body.AppendFormat("'{0}' is invalid, error was: {1}{2}", badUrl, badLinks[badUrl],
Environment.NewLine);
}

var smtp = new SmtpClient
{
    Host = "smtp.gmail.com",
    Port = 587,
    EnableSsl = true,
    DeliveryMethod = SmtpDeliveryMethod.Network,
    UseDefaultCredentials = false,
    Credentials = new NetworkCredential(fromAddress.Address, fromPassword)
};

using (var message = new MailMessage(fromAddress, toAddress)
{
    Subject = subject,
    Body = body.ToString()
})
{
    smtp.Send(message);

```



```
 */
```

```
SmtpClient client = new SmtpClient();
client.Port = 587;
client.Host = "smtp.gmail.com";
client.EnableSsl = true;
client.Timeout = 10000;
client.DeliveryMethod = SmtpDeliveryMethod.Network;
client.UseDefaultCredentials = false;
client.Credentials = new System.Net.NetworkCredential("group2forager@gmail.com",
"honeycomb");
```

```
StringBuilder body = new StringBuilder();
body.Append("The page ").Append(url).Append(" contains
").Append(badLinks.Count).Append(" invalid links:").Append(Environment.NewLine);
foreach (string badUrl in badLinks.Keys)
{
    body.AppendFormat("'{0}' is invalid, error was: {1}{2}", badUrl, badLinks[badUrl],
Environment.NewLine);
}
```

```
MailMessage mm = new MailMessage("donotreply@domain.com", "kkamdem@spsu.edu", "Forager
Report", body.ToString());
mm.BodyEncoding = UTF8Encoding.UTF8;
mm.DeliveryNotificationOptions = DeliveryNotificationOptions.OnFailure;
```

```
//add following code before smtpClient.Send()
```

```
client.EnableSsl = true;
```

```
ServicePointManager.ServerCertificateValidationCallback = delegate(object s,
X509Certificate certificate, X509Chain chain, SslPolicyErrors sslPolicyErrors) { return true; };
client.Send(mm);
}
```

```
public static string GetAppSetting(string key, string defVal, bool blnRequired, string
keyName)
{
    string value = ConfigurationManager.AppSettings[key] + "";
    if (value.Length == 0)
        value = defVal;
    if (value.Length == 0 && blnRequired)
        throw new Exception(string.Format("Missing application setting {0}!", keyName));
    return value;
}
```

```
public static string GetAppSetting(string key, string defVal)
{
```

```
    return GetAppSetting(key, defVal, false, string.Empty);
}
```

```
/// <summary>
/// Map all attributes of given element.
/// When same attribute appears more than once, only first value is taken.
/// </summary>
/// <param name="elementRawContents">Raw HTML of the element to parse</param>
/// <returns>Dictionary of attributes, key is the attribute name and value is its
value.</returns>
```

```
static public Dictionary<string, string> MapElementAttributes(string elementRawContents)
{
    //initialize result:
    Dictionary<string, string> attributes = new Dictionary<string, string>();

    //remove extra spaces:
    while (elementRawContents.IndexOf(" ") >= 0)
        elementRawContents = elementRawContents.Replace(" ", " ");

    //split by space:
    string[] parts = elementRawContents.Split(' ');

    //iterate over parts, analyzing each.
    foreach (string part in parts)
    {
        //look for equal sign..
        int eqSignIndex = part.IndexOf("=");
        if (eqSignIndex > 0)
        {
            //got attribute here. store name:
            string attName = part.Substring(0, eqSignIndex).ToLower();

            //maybe already exists?
            if (attributes.ContainsKey(attName))
                continue;

            //get value and add to result.
            string attValue = GetAttributeValue(part, eqSignIndex);
            attributes.Add(attName, attValue);
        }
    }

    //done.
    return attributes;
}
```

```
/// <summary>
/// Parse given content based on index of equal sign and return the value after the equal
```

```

sign.
    /// </summary>
    /// <param name="rawContent"></param>
    /// <param name="eqSignIndex"></param>
    /// <returns></returns>
    static public string GetAttributeValue(string rawContent, int eqSignIndex)
    {
        //maybe ends with it?
        if (eqSignIndex >= (rawContent.Length - 1))
            return string.Empty;

        //got something.. put aside:
        string value = rawContent.Substring(eqSignIndex + 1, rawContent.Length - eqSignIndex -
1);

        //need to calculate index of value start. default is right after equal sign.
        int startIndex = 0;
        int length = value.Length;

        //might start with quote, single or double:
        if ((value.StartsWith("'") && value.LastIndexOf("'") > 0) ||
            (value.StartsWith("\"") && value.LastIndexOf("\"") > 0))
        {
            startIndex = 1;
            length = (value.StartsWith("'") ? value.LastIndexOf("'") : value.LastIndexOf("\""))
- 1;
        }
        else
        {
            //might end with > or /> so handle this corner.
            if (value.EndsWith(">"))
                length -= 2;
            else if (value.EndsWith("/>"))
                length -= 1;
        }

        return value.Substring(startIndex, length);
    }

    /// <summary>
    /// This function will return all occurrences of the string between "start" and "end",
including both, inside the given raw contents.
    /// When no match is found, return empty array.
    /// </summary>
    /// <param name="rawContents">The string we want to analyze</param>
    /// <param name="start">Match start point, look for this string first</param>
    /// <param name="end">Match end point, return anything between start and this value</param>
    /// <returns>All occurrences between start and end</returns>

```

```

static public string[] GetBetween(string rawContents, string start, string end)
{
    //initialize return value.. can be any number of results
    List<string> results = new List<string>();

    //look for the first occurrence. Might not be found, in which case empty array is
returned.
    int indexStart = rawContents.IndexOf(start, 0,
StringComparison.CurrentCultureIgnoreCase);
    while (indexStart > 0)
    {
        //getting here means "start" was found. look for matching "end"
        int indexEnd = rawContents.IndexOf(end, indexStart + start.Length,
StringComparison.CurrentCultureIgnoreCase);
        if (indexEnd > indexStart)
        {
            //match found! add to results array and update the index:
            results.Add(rawContents.Substring(indexStart, indexEnd + end.Length -
indexStart));
            indexStart = indexEnd;
        }

        //search for next occurrence of "start" in the string: might reach end of string so
avoid error by limiting the count to string length.
        indexStart = rawContents.IndexOf(start, Math.Min(indexStart + 1, rawContents.Length
- 1), StringComparison.CurrentCultureIgnoreCase);
    }

    //done. return what we found:
    return results.ToArray();
}

/// <summary>
/// Little function to remove all HTML comments from given string.
/// </summary>
/// <param name="html">The raw HTML contents to be cleaned.</param>
static public void CleanHtmlComments(ref string html)
{
    int commentStart = html.IndexOf("<!--", 0);
    while (commentStart >= 0)
    {
        int commentEnd = html.IndexOf("-->", commentStart + 3);
        if (commentEnd > commentStart)
        {
            html = html.Remove(commentStart, commentEnd + 3 - commentStart);
            commentStart = html.IndexOf("<!--", 0);
            continue;
        }
    }
}

```

```

        commentStart = html.IndexOf("<!--", Math.Min(commentStart + 1, html.Length - 1));
    }
}

static public void WriteToFile(string fullUrl, string errorMessage)
{
    using (System.IO.StreamWriter file = new
System.IO.StreamWriter(@"C:\Users\karlloic\Documents\brokenlinks.txt", true))
    {
        file.WriteLine(fullUrl + " " + errorMessage);
    }

    //inserting result into database
    Database db = new Database();
    db.Insert(fullUrl, errorMessage);
}

}
}

```

### Database.cs

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using MySql.Data.MySqlClient;
using System.Windows.Forms;

namespace TestCrawler
{
    class Database
    {
        private MySqlConnection connection;
        private string server;
        private string database;
        private string uid;
        private string password;

        //Constructor
        public Database()
        {
            Initialize();
        }

        //Initialize values
        private void Initialize()
        {
            server = "localhost";
            database = "forager";
            uid = "root";
            password = "root";
            string connectionString;
            connectionString = "SERVER=" + server + ";" + "DATABASE=" +
                database + ";" + "UID=" + uid + ";" + "PASSWORD=" + password + ";";

            connection = new MySqlConnection(connectionString);
        }

        //open connection to database
        private bool OpenConnection()
        {
            try
            {
```

```

        connection.Open();
        return true;
    }
    catch (MySqlException e)
    {
        switch (e.Number)
        {
            case 0:
                MessageBox.Show ("Cannot connect to database. Contact Administrator");
                break;
            case 1045:
                MessageBox.Show("Invalid username or password. Please try again");
                break;
        }
        return false;
    }
}

//Close connection
private bool CloseConnection()
{
    try
    {
        connection.Close();
        return true;
    }
    catch (MySqlException e)
    {
        MessageBox.Show(e.Message);
        return false;
    }
}

//Insert statement
public void Insert(string link, string error)
{
    string query = "INSERT INTO report (link, error) VALUES(@link, @error)";

    //open connection
    if (this.OpenConnection() == true)
    {
        //create command and assign the query and connection from the constructor
        MySqlCommand cmd = new MySqlCommand(query, connection);

        //adding variables to insert statement
        cmd.Parameters.AddWithValue("@link", link);
        cmd.Parameters.AddWithValue("@error", error);
    }
}

```

```

        //Execute command
        cmd.ExecuteNonQuery();

        //close connection
        this.CloseConnection();
    }
}

//Update statement
public void Update()
{
}

//Delete statement
public void Delete()
{
}

//Select statement
/*public List <string> [] Select()
{
}

//Count statement
public int Count()
{
}*/

//Backup
public void Backup()
{
}

//Restore
public void Restore()
{
}
}
}

```