pylinkvalidator
Problem Statement : newAGEtech, Group H

Genevieve Okon (Okong), Abraham Omorogbe(Omorogoa),
Eric Le Forti(Leforte)

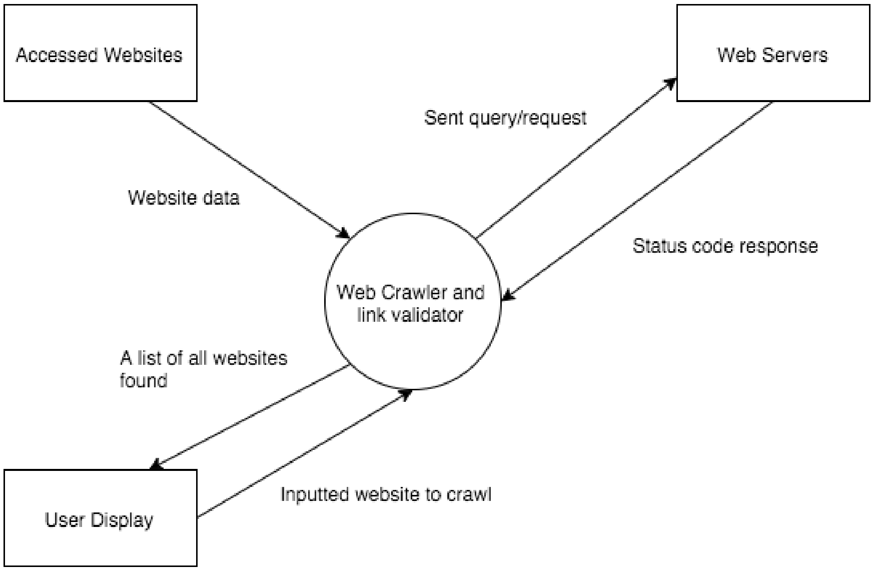
October 9, 2015

1 The Scope of the Work

1.1 The Current Situation

It is difficult often difficult for users of the internet to navigates through the web, it can be difficult to locate pertinent information using standard methods of search such as search engines. A web crawler and link validator will allow users quickly traverse the web and quickly and effectively locate information from different web pages and information about these web pages. A web crawler and link validator will increase users efficiency, it will be able to save money and resources due to users spending less time and effort surfing the web for information.

1.2 The Context of the Work



1.3 Work Partitioning

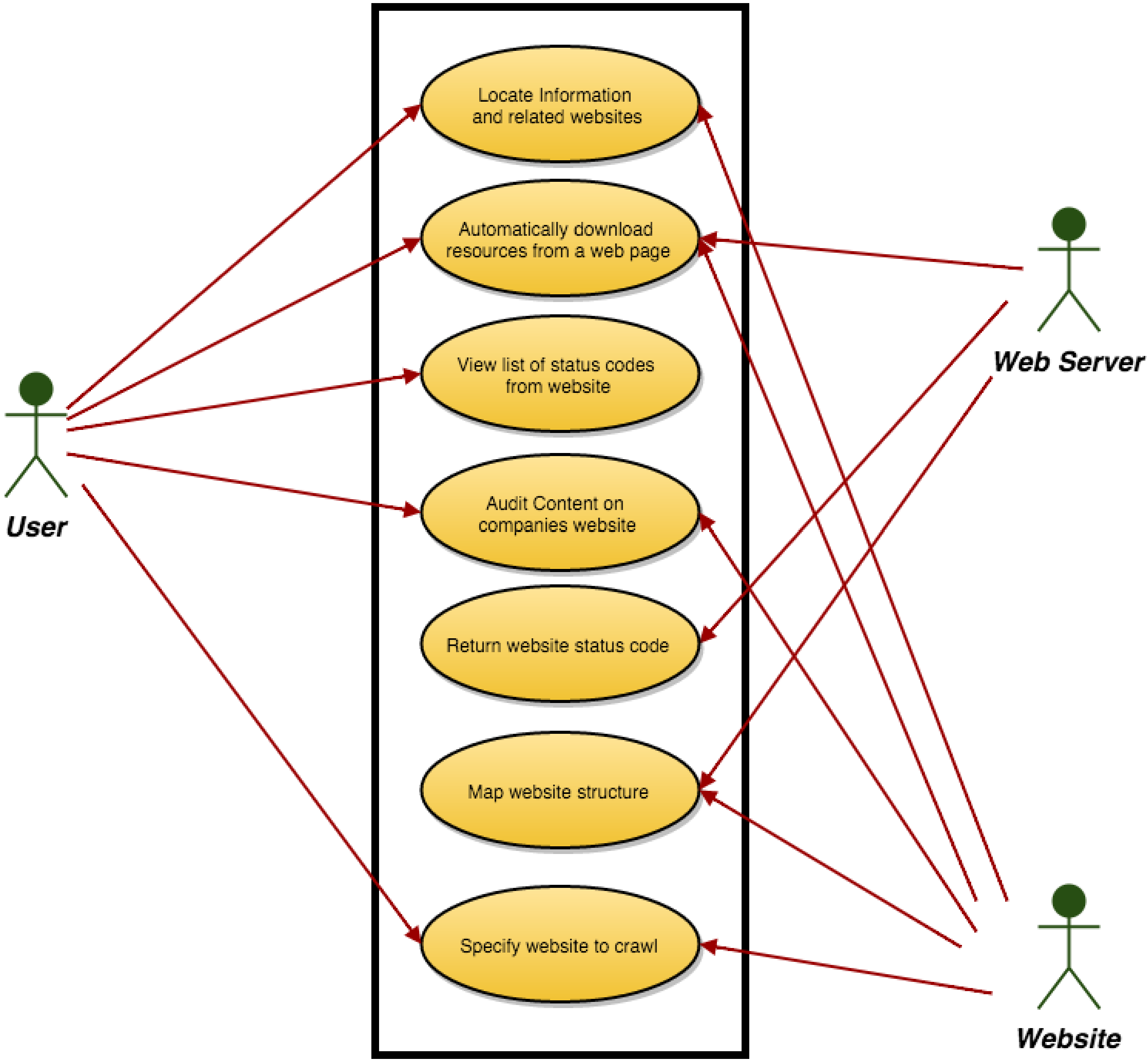
The table 1 is a Business Event List.

Event Name	Input and Output	Summary
User requests information about a website	Website start link to crawl and specific info gathering setting (In)	Have a starting point for crawler can gather resources from.
User requests information from a website	Website start link to crawl and specific status checking setting (In)	Have a starting point for crawler, it can continue to verify any links associate with initial website.
Traverse website	Links from initial retrieved website (In)	Has the abilities to reach website that are associated with starting link
List of crawled websites status code	A list of all the site that were visited (Out)	Has the ability to show the user the status codes from crawled websites, can verify if links are down.
List of information from website	All resources from a website (Out)	Has the ability to show the user all the information from website they have elected.

Table 1: Business Event List

2 The Scope of the Product

2.1 Product Boundary


Locate Information and related websites
Automatically download resources from a web page
View list of status codes from website
Audit Content on companies website
Return website status code
Map website structure
Specify website to crawl
User
Web Server
Website

2.2 Product Use Case List

- Locate Information and related websites
- Automatically download resources from a web page

- View list of status codes from website
- Specify website to crawl
- Audit Content on companies website
- Map website structure
- Return website status code

2.3 Individual Product Use Cases

1. Product Use Case Name: Locate Information and related websites
   Trigger: User requests a website to be processed
   Preconditions: User has to specify website that exists
   Interested Stakeholders:
   Actors: User, Website
   Outcome: If the website is a valid HTML page, the link on the page a verified and information is gathered (text-based), if the website is an invalid, an error is displayed.

2. Product Use Case Name: Automatically download resources from a web page.
   Trigger: User types in command to download data Preconditions: Web server has to have existing resources available, and user needs permission to the website.
   Interested Stakeholders:
   Actors: User, Web Server, Website
   Outcome: User has all the resources from the website (images, attached files etc.).

3. Product Use Case Name: View list of status codes from website Trigger: User types in command to show status code
   Preconditions: User must have Internet connection and must have entered a valid URL.
   Interested Stakeholders:
   Actors: User
   Outcome: User can see a list of all the status codes on websites associated with the initial site.

4. Product Use Case Name: Specify website to crawl
   Trigger: User types in a website to crawl
   Preconditions: User must have Internet connection and a valid website to crawl.
   Interested Stakeholders:
   Actors User, Website
   Outcome: The web crawler begins to traverse through website.

5. Product Use Case Name: Audit Content on companies website
   Trigger: User requests a list of all websites associated with the starting point.

Preconditions: User must enter a website with valid ¡a href=¿ ¡/a¿ tags
Interested Stakeholders:
Actors: User, Website
Outcome: The users can see a list of on the links and resources attached to every webpage associated with the specified domain.

6. Product Use Case Name: Map website structure
Trigger: User requests a structure of a website
Preconditions: Must have entered a valid URL/HTML.
Interested Stakeholders:
Actors: Web Server, Website
Outcome: User is shown the website file structure.

7. Product Use Case Name: Return website status code
Trigger: Web crawler sends request query to web server
Preconditions: Web server must respond to queries with status codes.
Interested Stakeholders:
Actors: Web Server
Outcome: The web server responds with correct status code.

3 Functional and Data Requirements

3.1 Functional Requirements

Requirement #: 1 **Requirement Type**: 9 **Event/Use case #**: 7
Description: The product shall return the status code of websites
Rationale: To allow the users verify the status of all the pages on a website. Checks if it is offline, Unauthorized etc.
Originator: Abraham Omorogbe
Fit Criterion: The web crawler displays an error with it reaches an website that is not online (Status Code: 200) and the status code related the error (Status Code: 400,500 etc.).
Customer Satisfaction: 5 **Customer Dissatisfaction**: 5
Priority: High **Conflicts**: None
Supporting Material: None
History: Created October 9, 2015 **Volere**

Requirement #: 2 **Requirement Type**: 9 **Event/Use case #**: 2
Description: The product shall download resources from a website
Rationale: To allow users quickly download all the HTML, images, CSS and other data related to

a website

Originator: Abraham Omorogbe

Fit Criterion: The download resources can be retrieved and used.

Customer Satisfaction: 5 **Customer Dissatisfaction**: 5

Priority: High **Conflicts**: None

Supporting Material: None

History: Created October 9, 2015 **Volere**

Requirement #: 3 **Requirement Type**: 9 **Event/Use case #**: 1,5

Description: The product shall have adjustable search max-depth

Rationale: To allow the user specific how many layers of the website, the user want to analyse. Helps with auditing and finding related sites.

Originator: Abraham Omorogbe

Fit Criterion: The web crawler only show top-level links when max-depth is at 0, and shows more levels of the website when depth ¿ 0

Customer Satisfaction: 5 **Customer Dissatisfaction**: 5

Priority: High **Conflicts**: None

Supporting Material: None

History: Created October 9, 2015 **Volere**

Requirement #: 4 **Requirement Type**: 9 **Event/Use case #**: 4

Description: The product shall always users enter a starting URL or local HTML page

Rationale: To allow the user crawl any website that is on or offline

Originator: Abraham Omorogbe

Fit Criterion: The crawler starts from the inputted website.

Customer Satisfaction: 5 **Customer Dissatisfaction**: 5

Priority: High **Conflicts**: None

Supporting Material: None

History: Created October 9, 2015 **Volere**

Requirement #: 5 **Requirement Type**: 9 **Event/Use case #**: 3,5

Description: The product shall display all a report of all found results and corresponding websites

Rationale: This is one of the main functionality of a web crawl, the application should be able to start at one sites and find related links. List will be used to gather info and check status codes

Originator: Abraham Omorogbe

Fit Criterion: The reports are accurate. All link that are reported as status code 404 are offline and all details of resource match source website.

Customer Satisfaction: 5 **Customer Dissatisfaction**: 5

Priority: High **Conflicts**: None

Supporting Material: None

History: Created October 9, 2015 **Volere**

Requirement #: 6 **Requirement Type**: 9 **Event/Use case #**: 5,6

Description: The product shall able to display websites structure

Rationale: Give to users a visually representation of the structure for entered website.

Originator: Abraham Omorogbe

Fit Criterion: The web structure matches the structure of inputted website.

Customer Satisfaction: 5 **Customer Dissatisfaction**: 5

Priority: High **Conflicts**: None

Supporting Material: None

History: Created October 9, 2015 **Volere**

Requirement #: 7 **Requirement Type**: 9 **Event/Use case #**: 1

Description: The product shall find specific information from specified websites and related sites

Rationale: To allow user enter details they are looking for, and the application can return data based on starting URL.

Originator: Abraham Omorogbe

Fit Criterion: Returned websites most be related with inputted website and information return must match what the user specified.

Customer Satisfaction: 5 **Customer Dissatisfaction**: 5

Priority: High **Conflicts**: None

Supporting Material: None

History: Created October 9, 2015 **Volere**

3.2 Data Requirements

- Valid HTML pages to be parsed and crawl through.
- The websites resources that the crawler can download and check for.