*Florida International University*

*School of Computing and Information Sciences*

Software Engineering Focus

Final Deliverable

Project Title:

Microdata and Algorithms 1.0

**Team Members: Michael A. Garrett, Christian Molto**

**Product Owner(s)**: DeEtta K. Mills

**Mentor(s)**: Robin Stubbs

**Instructor**: Masoud Sadjadi

The MIT License (MIT)

Copyright (c) *2016 Florida International University*

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

***Abstract***

*This document presents the information necessary to gain a good understanding of how to use the Microdata and Algorithms 1.0 genetic intensity data filter. The filter is intended for usage in the preprocessing stage of data mining applications that can be used for a variety of purposes. The idea to look for correlations in genetic intensity and time since death of human cadavers, as well as the idea to look for genes to construct a fungicide for laurel wilt were provided by the product owner, Dr. DeEtta Mills, who served as Director of the FIU Forensic DNA Profiling Facility in 2004. The filter has been constructed with a minimalist design to optimize performance, omitting the usage of objects and classes in favor of simpler data structures such as arrays. Due to this, the filter has not been greatly compartmentalized and does not have a particularly complex class structure, nor is it intended to be heavily modified after its initial deployment.*

**Table of Contents**

**Introduction** …………………………………………………………………………………………….....5

Current System ……….…………………………………………………………………………………………………..5

Purpose of New System …………………………………………………..……………………………. 5

**User Stories**

Implemented User Stories ………………………………………….…………………………………..6

**Project Plan**

Hardware and Software Resources ……………………………………………………………….… 7

Sprints Plan ……………………………………………………………..……………………………………………………….8

*Sprint 1*  …………………………………………………………………………………………………………………………...8

*Sprint 2*  …………………………………………………………………………………………………………………………... 8

*Sprint 3*  …………………………………………………………………………………………………………………………... 8

*Sprint 4*  …………………………………………………………………………………………………………………………... 8

*Sprint 5*  …………………………………………………………………………………………………………………………... 8

*Sprint 6*  …………………………………………………………………………………………………………………………... 9

**System Design**

Deployment Diagram …………………………………………………………………………………………………….…... 10

**Appendix**  ……………………………………………………………………………………………………………………………….11

Appendix A - UML Diagrams ………………………………………………………………………………………………. 12

*Static UML Diagrams*  ……………………………………………………………………………………………………….12

*Dynamic UML Diagrams*  …………………………………………………………………………………………………..12

Appendix B - User Interface Design …………………………………………………………………………….…... 13

Appendix C - Sprint Review Reports …………………………………………………………………………...…… 19

# 

# Introduction

## Current System

The Biology Department was used to Microsoft Access to query their data but it did have shortcomings. The Microsoft Access did not have a range filter that can help the team to find genes that were specifically desired. The current system also could not differentiate between one entity from another that is to say that the data points belonging to one cadaver were to be separate from data points of another cadaver. There was a filter for other inputs like gene name, categories, lineage and even organism names. However, the current system could not take them all at once. Other features not present in the current system were limiting the number of present data points, separation of different entities, deleting unique items and omitting data points that are not within a user-provided range.

## Purpose of New System

The program was designed from the ground up to give the department a system that had everything that was desired and agreed upon. The system has a range filter that allows the user to have a range to work with the data and reduce it to at a more workable level. There are multiple name filters to allow the user to input gene name, lineage, organism name and categories with no worry of case sensitivity. The program can also allow you as the user the option to put a limit on number of data points to further reduce the dataset so that the user can see for example rows of data that have only three data points. The user can even type how many rows that appear a certain amount of times to delete any unique item. The program can even replace data points that are outside the range the user specified with an asterisk and insert empty points with a letter.

# User Stories

The following section provides the detailed user stories that were implemented in this iteration of the Microdata and Algorithms 1.0 project. These user stories served as the basis for the implementation of the project’s features. This section also shows the user stories that are to be considered for future development.

## Implemented User Stories

The development team have implemented user story #667 which is the filter for organism and species displaying rows only of the organism and species you specified. Another user story #668 which filters data by using a minimum and maximum range finder. User story #671 which the time data point threshold designed to have a filter to only display rows of data containing the number of data points the user have inputted. The next user story #679 which filter other categories like separating columns that represent two different cadavers. The following user story #683 which deals with the filtering of gene names that appear a user defined amount of times and they are promptly deleted. The user story #687 sets the other filters already implemented and tells the program which has the priority to go first, second and so on. The final user story #688 sets inputs for the filters in a case insensitivity way to avoid any problems with inputs that required precise cases.

# Project Plan

This section describes the planning that went into the realization of this project. This project incorporated the agile development techniques and as such required the sprints to be planned. These sprint plannings are detailed in the section. This section also describes the components, both software and hardware, chosen for this project.

## Hardware and Software Resources

The following is a list of all hardware and software resources that were used in this project:

* **NetBeans 8.1**
* **Java 8**

## 

## 

## Sprints Plan

### Sprint 1

The first sprint was to set up the development environment by using R application. The next thing was to find rules for association and garner user stories. Found out the R-php environment was incapabable to project use and Hurricane Irma forced the development team to use java based application instead.

### Sprint 2

The second sprint was dealing with implementing Java application for project use which went well. Intensity filter worked well. Attempted to use Weka in Java for data mining techniques and use of random Tree algorithm. There was a problem as the memory heap was an error that occured in this stage of the project.

### Sprint 3

The third sprint was spent working on the intensity filter and the other categories filter. We considered using MicroStrategy Desktop but it proved to be insufficient. The application showed the team that the genes that have some similarity with each other and differences with each other. We also gained new user stories from the project owner for extra features.

### Sprint 4

The fourth sprint was spent to work on the organism and gene name filter for the project. The Superkingdom filter was also implement as part of other categories filters along with intensity range filter was also integrated into the project. The intensity data point filter was also designed and implemented at this stage as well.

### Sprint 5

The fifth sprint was dealing with adding additional features to intensity filter as well as remove duplicate rows of data printing to output file and working on the column operations which will separate the two data entities from one another. Filter priority component was worked on and completed in this sprint. The case insensitivity part of the project was worked on this sprint.

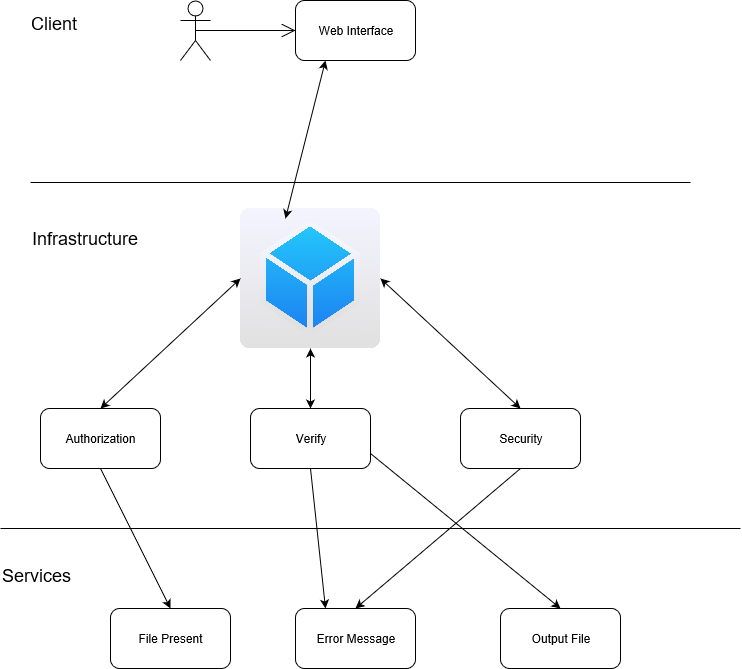
### Sprint 6

The sixth sprint was the one that team finished the project. The ignore case filter was completed after a week’s work and integrated into the program. The team needed test data from project owner which was received and send results to the project owner to see if the results are what the project owner wanted. The preprocessing against the human microbiome user story was omitted to minimize overhead, as the human microbiome was not complete at the time of this writing and creating a system to read it from the internet would slow system performance to a level that does not meet the user-specified requirements.

# System Design

This section contains information on the design decisions that went into this project. The architecture patterns are outlined and explained. The entire system is shown in a package diagram and the subsystems are explained. Finally, the design patterns used in the project are discussed.

## Deployment Diagram

****

# Appendix

## Appendix A - UML Diagrams

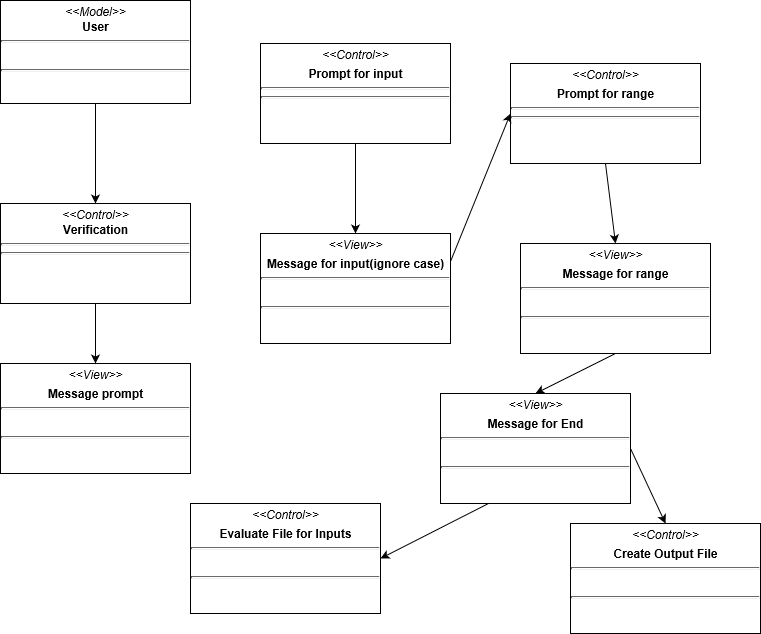
## 

Fig. 1. General Use Case.

## 

## Fig. 2. General Sequence Diagram.

## Appendix B - User Interface Design



## Appendix C - Sprint Review Reports

## Sprint 1 Review Meeting Minutes

Review Meeting Minutes

Attendees: Michael Garrett, Christian Molto, DeEtta Kay Mills, Robin Stubbs

Start time: 4:30 pm

End time: 6:00 pm

After a show and tell presentation, the implementation of the following user stories were accepted by the product owners: All.

* 667 Filter by Organism/Species
* 668 Filter by Intensity
* 669 Correlation Discovery
* 670 Intensity Depletion Discovery
* 671 Time Data Point Threshold
* 672 Development Environment

The following ones were rejected and moved back to the product backlog to be assigned to a future sprint at a future Sprint Planning meeting.

* 667 Filter by Organism/Species
* 668 Filter by Intensity
* 669 Correlation Discovery
* 670 Intensity Depletion Discovery
* 671 Time Data Point Threshold
* How this should be reflected on the user story definition in Mingle:
  + The user stories can be done just not in this sprint due to environmental issues.

## Sprint 2 Review Meeting Minutes

Review Meeting Minutes

Attendees: Michael Garrett, Christian Molto, DeEtta Kay Mills, Robin Stubbs

Start time: 4:30 pm

End time: 6:00 pm

After a show and tell presentation, the implementation of the following user stories were accepted by the product owners: All.

* 668 Filter by Intensity
* 669 Correlation Discovery
* 671 Time Data Point Threshold
* 672 Development Environment

The following ones were rejected and moved back to the product backlog to be assigned to a future sprint at a future Sprint Planning meeting.

* 670 Intensity Depletion Discovery
* 679 Filter other categories
* 680 Filter out gene names appearing <x times
* 681 Cadaver preprocessing against human genome
* How this should be reflected on the user story definition in Mingle:
  + These user stories have been added to the mingle Release 1 > Product Backlog. It was decided through discussing with product owner and with input on current user stories what to use on the project. More details will come during future sprint meetings.

## Sprint 3 Review Meeting Minutes

Review Meeting Minutes

Attendees: Michael Garrett, Christian Molto, DeEtta Kay Mills, Robin Stubbs

Start time: 4:30 pm

End time: 6:00 pm

After a show and tell presentation, the implementation of the following user stories were accepted by the product owners: All.

* 683 Filter out gene names appearing <x times

The following ones were rejected and moved back to the product backlog to be assigned to a future sprint at a future Sprint Planning meeting.

* 670 Intensity Depletion Discovery
* 679 Filter other categories
* 680 Filter out gene names appearing <x times
* 681 Cadaver preprocessing against human genome
* How this should be reflected on the user story definition in Mingle:
  + The user stories will be moved to their appropriate sprints.

## Sprint 4 Review Meeting Minutes

Review Meeting Minutes

Attendees: Michael Garrett, Christian Molto, DeEtta Kay Mills, Robin Stubbs

Start time: 6:00 pm

End time: 6:02 pm

After a show and tell presentation, the implementation of the following user stories were accepted by the product owners: All.

* 683 Filter Out Gene Names Appearing <x Times
* 667 Filter By Organism/Species

The following ones were rejected and moved back to the product backlog to be assigned to a future sprint at a future Sprint Planning meeting.

* 670 Create Intensity Depletion Discovery
* 671 Create Time Data Point Threshold
* 679 Filter Other Categories
* How this should be reflected on the user story definition in Mingle:
  + The user stories will be moved to their appropriate sprints.

## Sprint 5 Review Meeting Minutes

Review Meeting Minutes

Attendees: Michael Garrett, Christian Molto, DeEtta Kay Mills, Robin Stubbs

Start time: 6:26 pm

End time: 6:30 pm

After a show and tell presentation, the implementation of the following user stories were accepted by the product owners: All.

* 670 Create Intensity Depletion Discovery
* 671 Create Time Data Point Threshold
* 679 Filter Other Categories

The following ones were rejected and moved back to the product backlog to be assigned to a future sprint at a future Sprint Planning meeting.

* 681 Cadaver preprocessing against human microbiome
* How this should be reflected on the user story definition in Mingle:
  + The user stories will be refined and set to appropriate sprints.