Your project title: Which can be longer when it is displayed on the front page of the document

Report Name	Outline Project Specification
Author (User Id)	Adam Connah (aoc9)
Supervisor	Georgios Gkoutos (geg18)
Module	CS39440
Scheme code	H621
Scheme title	Internet Computing
Date	October 21, 2012
Revision	1
Status	Release

1. Project description

The project I am undertaking for my final year is to create an RDF store that can be used to provide a fast analysis of the similarity between different phenotypes in animals. This knowledge can then be used to aid our understanding of the function of genes, by systematically analyzing the phenotypic outcome of their mutation and ultimately facilitate our ability to prioritize causative genes for rare and orphan diseases by enabling the comparison of the phenotypic similarities between experimental data and human clinical signs and symptoms.

Using an RDF store is key to the project as it is a decentralized system, and hence it facilitates the retrieval and analysis of data that resides in a variety and diverse sources. Such a store will generate a large resource that can be accessed quickly and efficiently providing biomedical researchers and bioinformaticians alike a valuable resource and tools for their analysis.

The system needs to be use friendly, with a very easy to use GUI (graphical user interface), as well as providing the ability for the user to search common keywords and retrieve and present results in a comprehensive and easily understood manner. This will require a toolkit to deliver a professional and easy to use GUI, as well as using SPARQL (a query language for RDF), to achieve the advanced retrieval capabilities.

2. Work to be tackled

At first I will need to gain a good understanding of the new technologies I will be using, as I have not used many of them before. This will require a lot of research into the different areas, and doing some spike work to gain a better grasp of the coding parts. This will include testing out different toolkits, such as the Google web toolkit, to see which is best suited for what I want to achieve. I will also need to decide upon a Data Management system that is capable of dealing with RDF triple stores, and become accustomed to interacting with it. I have decided to code in PHP for the project, and so throughout the project I fully expect to have to research new methods and techniques, and this will help to build upon my current knowledge of PHP.

In terms of large parts of the project, I foresee a few parts of the project which will be main points to focus on. The first of these will be the GUI. This will require a lot of planning before any implementation. I will need to look at similar existing systems, and see which parts of them are a common theme. This will help not confuse the user and will keep a sense of consistency so they have a better experience with the application. I will also think of ways to add better functionality to the current systems, and ways that certain aspects could be portrayed better. This includes the representation of the results, which needs to show the relevant information straight away, and be sortable to the user's preference.

Setting up the RDF store will require communicating with various different data stores, and using the standards already in place to be able to present the data properly. This is definitely the most desirable part of RDF, and as a result it allows us to link to any data that we are interested in, providing it is a triple store, and that data can even use its own “vocabulary”, which is like a schema for that set of data. RDF focuses on the relationship between data, to try and give meaning to it. This makes it very “human friendly” as it gives us the option of doing more than just simply presenting it. RDF uses logic to present a fact, and applies this to create knowledge of real things. This closely relates to a technology I will need to investigate, called JSON. * what is json* This is effectively a Javascript application that runs in the browser and communicate with a server using XML.

Another technology I will need to investigate and research, is the SPARQL language. This is a similar concept to languages like MySql, which allows users to query databases. In this instance the SPARQL language will be used to query the RDF store. I don’t think this part of the project will be too difficult, as it seems SPARQL has similarities to query languages I have used before.

Throughout the project I will need to develop a basic understanding of the data I am working with. This will help me to better understand how to present it, and how to interact with it. It will also give me a chance to see how Computers can be used to help in aspects of life, you would not commonly associate them with, in this instance, Biology.

Obviously the project will need some form of testing to occur at different points throughout. I think it will be a good idea for lots of testing to occur throughout, so that I am confident each small part is doing the correct job. This will give me more confidence in the application when it is all put together. Key aspects to look at will be:

- Security - To make sure that the application cannot be broken by malicious code or queries.
- Functionality – To make sure the application behaves as expected.
- Performance – To make sure that the application is responsive and efficient.

3. Project deliverables

Initial project plan – This is very similar to the Outline Project Specification, but this is for my Dissertation supervisor, to show them that I have grasped an understanding of what needs to be achieved.

Progress report – This will include a description of work undertaken so far. It may also include any differing routes taken with the project, or mention any problems not foreseen at the start of the project that could have an effect on the outcome.

Mid project demonstration – This provides an assessment of the work undertaken before christmas/early in semester 2, and to focus on what will need to happen next to achieve the preferred result.

Final report – This report will focus on what I have achieved throughout the project. Including, but not limited to the processes that have taken place, the planning, design, building, and testing. It will also talk about the things I have learnt during the project, as well as providing a critical evaluation of my own performance.

Final demonstration – This is where I get a chance to show the application to the Project supervisor and other bodies.

4. Initial bibliography

[1] Introduction to RDF. www.rdfabout.com/quickintro.xpd/ Accessed October 2012

[2] RDF Store: http://virtuoso.openlinksw.com/ Accessed October 2012

[3] Google web toolkit with PHP. Http://www.gwtphp.com Accessed October 2012

[4] Tutorial on Google Web Toolkit: https://sites.google.com/site/angelhurtado/tutorialgwt2 Approx before 2009. Accessed October 2012

[5] Yahoo user interface library. http://yuilibrary.com/projects/phploader/

[6] PDF Provided by Georgios Gkoutos. Introduction to genetic research and the use of Ontologies. *Vision. Not dated.*

[7] Biology technical terms and phrases provided by Georgios Gkoutos.