Honours Project Proposal

Zamani Data Archive

Revised: 18 June 2014

Proposal compiled by:
Michael Ferguson (FRGMIC005)
frgmic005@myuct.ac.za

Jay Benson (BNSJAY001)
bnsjay001@myuct.ac.za

Supervisor:
Hussein Suleman
hussein@cs.uct.ac.za

Table of Contents

1. Project Description

The Zamani Project was initiated to create a permanent metrically accurate record of important heritage sites in Africa and the Middle East as well as increase the international awareness of these sites. The purpose of this preservation is fourfold in that the data has been collected for education, research, restoration and conservation purposes [3].

Thus far, the Zamani team has documented approximately 40 sites in 12 countries in Africa and the Middle East, amounting to close to one hundred three-dimensional (3D) models of individual structures. The dataset collected by the team contains approximately 44TB including backups and copies. The data comprises of different forms such as 3D models, plans, videos, panorama imagery and Geographic Information Systems (GIS) [2].

The project currently requires an archive for its large collection of geospatial data. The archive will require an archival storage management system and a way of presenting the data to end users.

2. Problem Statement

The Zamani Project is currently run by four team members; Heinz Ruther, Roshan Bhurtha, Ralph Schroeder and Stephen Wessels. With much of their time dedicated to field work and the collection of data there is a need for the creation of a digital archive to store it. The Zamani data set currently consists of partially structured, geospatial data.

The Zamani team require an archive for their large geospatial data collection of cultural heritage sites. The problem consists of two major aspects, namely: archival storage management and end-user presentation. The archive will include a Metadata Management Tool to ensure data has complete metadata before it is added to the archive.

Researchers and the general public require access to the data as one of the primary objectives of the Zamani Project is creating awareness for the sites. The archive will feature a Public Portal that will enable end-users to search, view, request and download data from the archive. The transmission of the files must be considered for efficiency due to the large file size found in Zamani's dataset.

Currently the data collected by the Zamani Project is not complete in terms of having full descriptive metadata. Metadata is limited to the sites that were shared with Aluka and, additionally, the existing metadata is inconsistent [5]. The partial automation of the missing metadata is thus an important problem that needs to be addressed.

Additionally, access to the data collected by the Zamani Project must be controlled so that the data is used for its intended purpose and not exploited for commercial reasons.

A diagram that outlines the user interactions with the proposed system can be seen below.

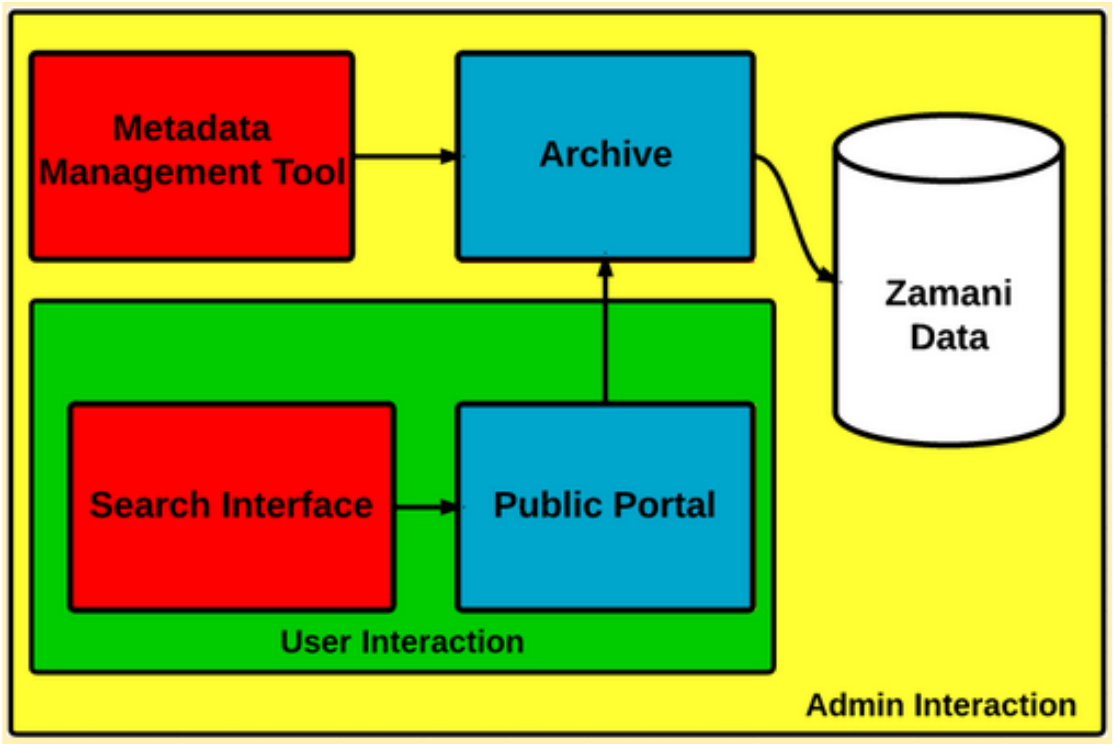


Figure 1: Overview of the proposed system

3. Procedures and Methods

Implementation Strategy

The implementation strategy adopted will be the Agile Programming Methodology. It will consist of daily stand-up meetings wherein group member progress and current issues will be discussed and analysed. Sprint cycles that are approximately one week long will be used during the development phase. Sprint planning sessions will be held at the end of each sprint cycle, with the addition of regular meetings with the project supervisor.

Development Platform

DSpace

The core archiving tool used in this project will be Dspace. DSpace is a repository software framework that is open source and typically used in the creation of public access repositories for academic institution's digital content. It provides a digital archiving system that focuses on the long term preservation and access of digital content [8]. DSpace consists of a set of java web applications and utility programs that maintain the relationship between the data and metadata. There are various interfaces that the web applications provide such as: search, access, administration and ingestion. The maintenance of the data occurs on a file or storage system. Metadata is stored within a relational database, DSpace supports two such databases, namely: PostgreSQL and Oracle [9].

DSpace makes use of a qualified Dublin Core Metadata standard, wherein only three fields are required: title, submission date and language. All additional fields are optional fields [8].

Additionally, DSpace offers search interface functionality. This includes a number of different capabilities including; faceted searching, sorting by relevance and advanced search [9]. There are two web interfaces that DSpace supports, namely JSPUI and XMLUI (Manakin). JSPUI makes use of JSP and the Java Servlet API while XMLUI uses XML and XSLT [10].

Design Features

Metadata Management Tool

A Metadata Management Tool will be included in the proposed system to ensure that descriptive data is preserved. Additionally it will aid in the creation of a rigid structure for the archive.

Automation

The Metadata Management Tool will follow the Dublin Core Metadata standard as it is the ingest format used by DSpace [8]. Additionally, the tool will allow for partial automation of some of the metadata fields. The extent to which metadata is automated will depend on the descriptive data contained within the dataset. The output of this tool will be an XML file containing the metadata for the data set being managed.

Layered Metadata

The inclusion of an additional metadata layer may be needed if any other software packages used require specific fields or formats for the metadata. The metadata automation process outputs files in the ingest format used by DSpace, thus it is unlikely that there will be a need for a metadata layer [8]. However, an additional metadata layer could be used to enforce

relationships between data contained in the Zamani data set. Should an additional metadata layer be generated, it will inherit the characteristics of the existing metadata where possible.

Search Interface

A Search Interface will be created to provide navigation through the data set and to provide a representation of the data. This interface will seek to improve on issues present in Aluka's text-based search interface [4].

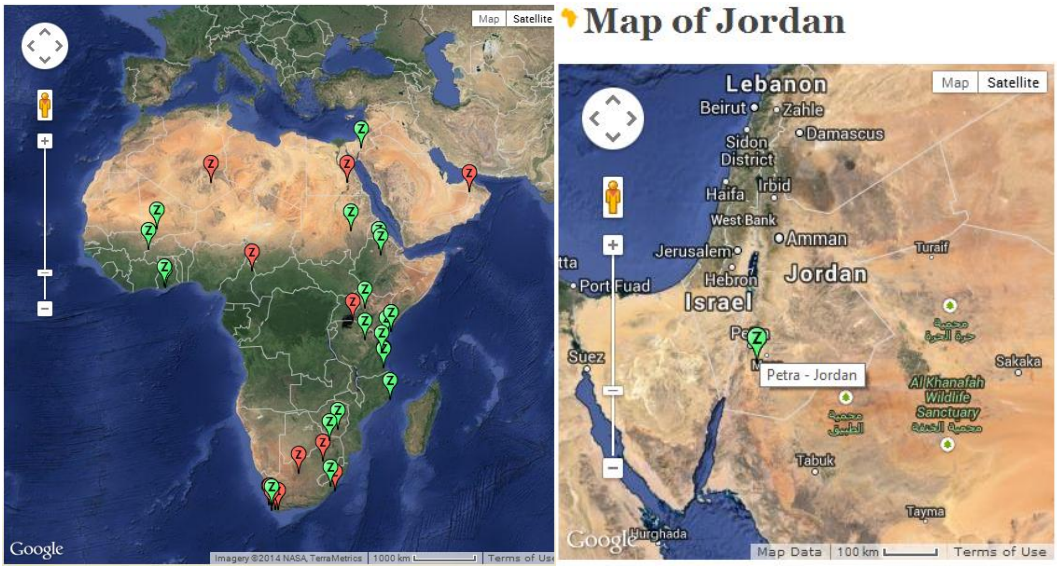
Text-based and faceted search will be accomplished with the use of DSpace Discovery [11], an addon for DSpace's XMLUI (Manakin) web interface [10] . It succeeds in replacing typical DSpace search and browsing with that of SOLR's [11].

Text-based search

Conventional text-based search will be present in the interface. This will allow the user to perform a number of different queries based on the descriptive details of heritage sites, such as their location or coordinates. Search results will be ordered by relevance to the search query with the most relevant results being displayed first.

Google Map Integration

The Zamani Project [1] currently makes use of Google Map Integration on their website, as can be seen below (www.zamaniproject.org). It consists of an overlay of existing sites and will be used as an additional method to navigate through the data in the archive.



Faceted searching

In order to reduce the time a user spends searching through multiple results returned by a search query, they will be allowed to filter their results on a number of facets, including file type, location and date.

Public Portal

The public portal will serve as the central point of interaction with the Zamani Archive. It will be responsible for requesting data and distributing it where requests are permitted. The portal will consists of a number of components, namely a search interface, data request interface and video streaming functionality.

Video Streaming Service

It has been decided that a streaming service will be more appropriate for video content on the portal due to the large file size of the videos currently contained in Zamani's data set.

Distribution

The public portal should also provide the user with a means to request and receive files. This will require a user to register before they can make a request for data in the archive. The portal will also notify an administrator when a request is made and allow them to set the relevant permissions for the data in response. All file transfers will use HPN-SSH, which is a high performance SSH/SCP implementation released as a patch for OpenSSH [6].

User Permissions

In order to ensure that Zamani's data policies are adhered to, a User Permission System will be integrated into the Public Portal. The key stakeholders of the project are the Zamani Team, researchers, and cultural heritage enthusiasts. Due to the different user requirements of the specified stakeholders, a permission system will be required. This system will provide access control for the Zamani data set. This will ensure that users only have access to data that they have requested and have subsequently been permitted access to by an administrator.

The Zamani Team will act as the administrators of the system. All stakeholders will have access to a representation of the data as one of the primary objective of the Zamani Project is to create awareness of cultural heritage sites. The platform allows the administrators control over READ, WRITE, ADD and DELETE access of digital items [14]. Additionally, the system allows for the licensing of digital items.

Research Data Archive

The primary component of this project is the data archive itself. The Metadata Management Tool will be used to ensure complete metadata is present for all the data in the Zamani data set. Once complete metadata is available, the archive will be responsible for structuring the data

and ensuring that any existing links are maintained. The archive will be responsible for a variety of administrative tasks, such as adding, removing and updating heritage sites.

DSpace will be the platform used to create the archive.The built-in functionality in DSpace uses manual submission tools which will not be used unless a small dataset needs to be processed [7]. In order to deal with the large amount of data held by the Zamani Project this project will make use of the import/export functionality that DSpace provides. It provides tools namely, the Item Importer and Exporter, created for mass digitization [12]. This is done through a process whereby metadata is standardised for digital items and put into the archive [7].

A Postgresql database will be used for the system. This relational database has been chosen for a number of reasons. It is an open source alternative that already has existing support for DSpace and provides a number of tools that can be easily integrated with the database to aid in visualisation and the processing of data. Postgresql also offers a performance boost and more functionality for spatial data when using the PostGIS database extender [15].

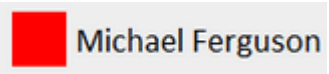
Expected Challenges

- Organisation of the complex data that the Zamani Project has collected.
- Partial metadata automation.
- Streaming of large videos.
- Secure transmission of large data files.
- Search interface for geospatial data.
- Integration challenges between components.

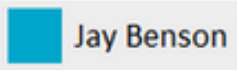
Work Allocation

The project will be split up into 4 components. Each of the group members will be allocated to two of the components, which will include doing the research and the implementation for both components. Additionally, comprehensive testing of this system will be performed by both team members.

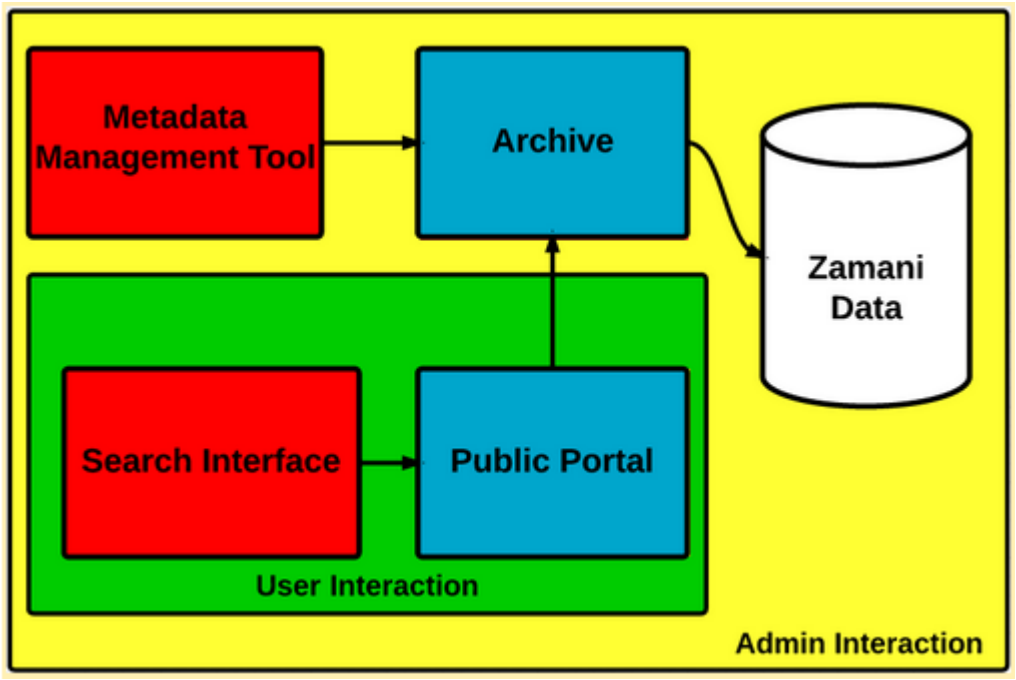
The components allocated to each group member are as follows:



- Metadata Management Tool
- Search Interface



- Public Portal
- Archive



Testing Plan

The testing plan will consist of peer review and User Acceptance Testing.

Before a component is integrated into the system, it should be tested by the group member responsible for the component. Once a component has been integrated, it should then be tested by the other team member using a series of test scenarios. This is done to ensure that the component is of a high standard and is bug free.

User Acceptance Testing will be conducted in order to determine whether the requirements of the system have been met. It will be conducted by both group members, with the participants being the Zamani team. The reason for using the clients of this project as the participants for user testing is fundamental, as it will ensure that the security and level of user interaction with the system is designed to their desired requirements. Ideally, any bugs that are present will become apparent during testing as well as any user interaction requirements that have not been met. Additionally it can be used to gauge the level of user interaction with the developed system as the Zamani team are all researchers, which is the typical user group.

As this project entails user testing, it is important that ethical clearances are obtained and that all participants provide their consent before participating. Ethical, Professional and Legal Issues are addressed in the section below.

4. Ethical, Professional and Legal Issues

Ethical Issues

Before the research is continued, it is essential that ethical clearance is obtained. Additionally, before any user testing it must be ensured that participants have provided consent to participate.

As previously mentioned, the user testing that has been proposed will be conducted using the Zamani team. Ethical clearance from the Research Ethics Committee will be needed in order to conduct such tests. It is fundamental that we apply for ethical clearance early to avoid any unforeseen delays which could lead to the delay of user testing.

Legal Issues

Precautions will be taken to ensure that no copyright of proprietary research or software is infringed upon. Any software used will be primarily open source and will be acknowledged. Any proprietary software that is used will be accompanied by legally obtained software licences.

The intellectual property of the developed system will be property of The University of Cape Town and the project team (Michael Ferguson, Jay Benson).

Utmost precaution must be taken to ensure that Zamani's data policy's are not infringed on. Zamani's data may only be used for education, research and, with special permission, for the restoration and conservation of cultural heritage sites. Under no circumstances may the data be used for commercial purposes, therefore the system's security needs to be ensured [3].

All participants gathered for user testing will be required to fill out a consent form and will be notified that the results will be kept confidential.

5. Related Research

When considering the implementation of this project the following research has been reviewed. It will serve as background information and aid with the understanding and development of the main system components.

Aluka Archive

Aluka [5] is a collaborative international programme that is aimed at creating an online digital repository about Africa (http://www.aluka.org). It archives non-spatial data such as books and

scientific papers in digital form as well as presents its spatial data in the form of GIS, Spatial Information Systems (SIS), 3D models, elevation views, ground plans, sections and panoramas. It makes use of various technologies to acquire such data, namely: photogrammetry, laser scanning, remote sensing, image processing, conventional surveying, GIS and CAD. Additionally, all of the spatial data is associated with metadata about the survey details. Aluka had a total of nineteen sites that had been documented by the end of 2009.

Repository Platform - National Library of Finland

The National Library of Finland have proposed DSpace as an appropriate repository platform for the management of geospatial data of cultural heritage sites [7]. However, it is made clear that DSpace should not be viewed as a standalone system, but rather as a component of a larger system infrastructure. As it is essential to have well planned processes for the dissemination of metadata and data between systems when digitization is occurring at a mass scale. Additionally, the National Library of Finland proposed that DSpace is generic enough, and can be used in a variety of use cases. This is due to its ability to be linked to other systems through technical interfaces, and by being open source, it is easily modifiable.

6. Anticipated Outcomes

Software

This project should create not only a research data archive but instead a complete framework for the archiving, linking, representation, distribution and replication of the data collected by the Zamani Project.

Expected Results

Metadata Management

Metadata should exist for all the data within the archive that is created. Where metadata is currently missing, it will be created using the Metadata Management Tool that will aid the process through automation where possible.

Search Interface

The search interface should allow any user of the archive to easily search and navigate through a representation of the data found in the archive. This will include conventional text-based searching, Google Map searching and faceted searching.

Public Portal

When the public portal is integrated with the search interface, a user will be able to view a representation of the data found on the archive. A video streaming service will be made available for any video resources found on the archive.

The portal will make provision for requesting files and distributing them to users. Due to the large file size, the download process should be stable and as fast as possible.

Research Data Archive

A well structured archive will be created. The archive will have an administrative interface. This will allow the administrator to perform maintenance on the archive. Maintenance tasks include: ingestion of data, metadata management, permission and access management as well as standard CRUD functionality for the data found within the archive.

Impact of Project

Preservation of heritage sites throughout the world has been the goal of enthusiasts and researchers alike. Spatial data acquisition repositories such as Zamani have proved [5] to be invaluable in facilitating quantitative analysis and planning of the preservation of heritage sites. Aside from providing permanent records of cultural heritage for future generations, it has potential to aid in providing practical quantitative planning of the conservation and restoration of such sites. Finally, with regard to education and tourism, a project such as Zamani serves as a means to publicize African heritage.

Finding an effective way to archive the large collection of interlinked geospatial data held by the Zamani Project is a significant impact of the project, as the team currently have no tools to perform such tasks. Additionally, this project serves to explore this field in an attempt to provide future preservation projects with an efficient model to work from.

Key Success Factors

The most vital success factor of this project is an overall positive response from the Zamani Team. As the nature of the project is envisioned to be long term, any changes or additions required by the Zamani may be viewed as possibilities for future development and not project failure.

The effectiveness of managing the interlinked nature of the information and retaining structure within the archive, will be used to measure the success of the tools used to create the Zamani Archive. The system will require the right combination of components for this to be possible.

Additionally the success of the project will be dependent on:

- High adoption rate by the public.

- Sending the large files in an efficient manner.
- Completing the project on time.

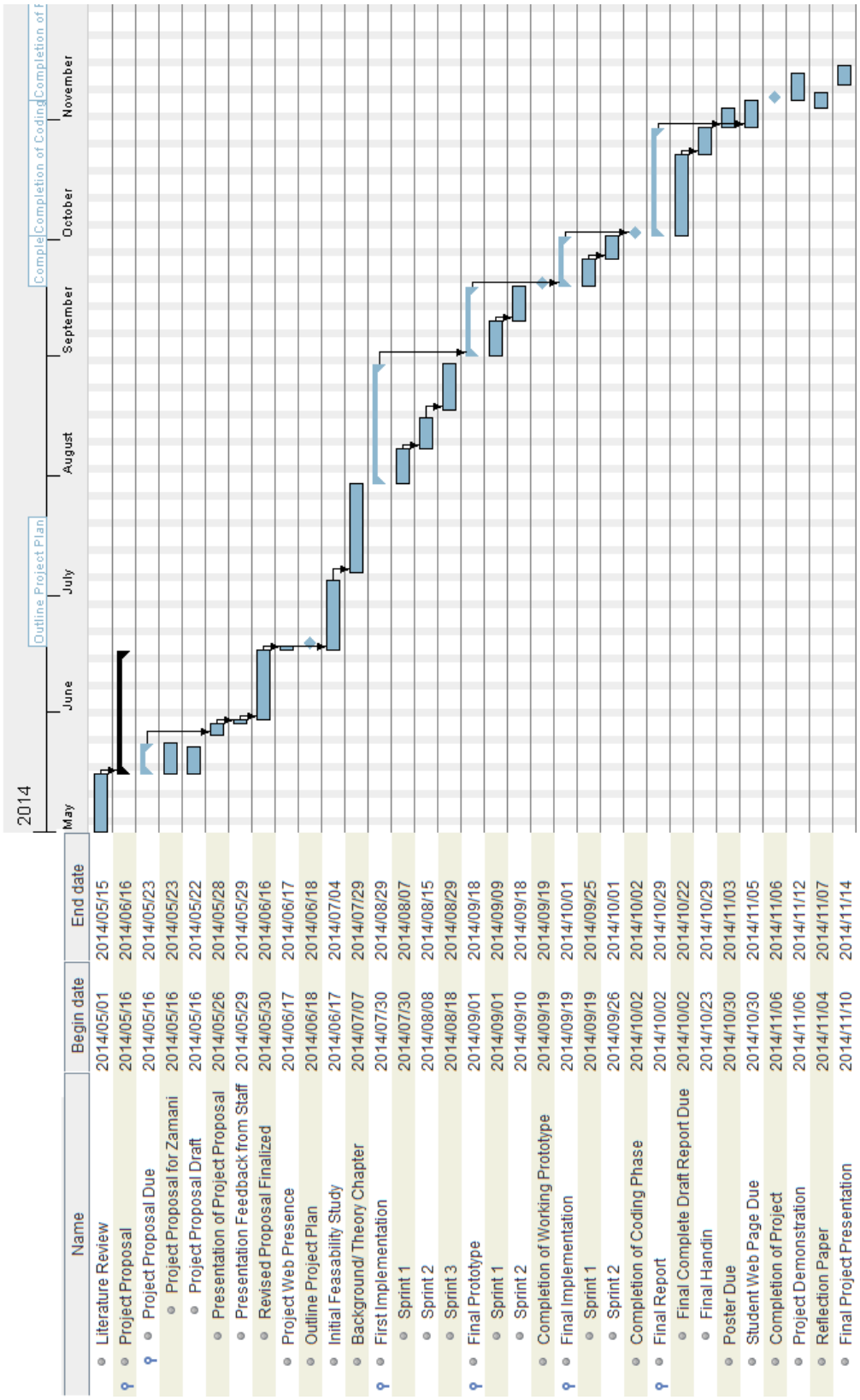
7. Project Plan

Risk and Mitigation

Risk	Risk Type	Details	Impact	Likelihood	Mitigation Strategy
Lack of user involvement in defining requirements	Market and People	The product may not fulfill the requirements of its users.	Medium	Low to Medium	Regular meetings with the Zamani team.
Lack of adoption	Market and People	Researchers do not actively and willingly participate in using the data.	High	Low	Active engagement with the research community, ensuring that the proof of concept is fully implemented.
Data Copyright Infringement	Legal and Regulatory	Data is not used for its intended purpose (e.g. 3D models used as a game model).	High	Low	- A permission system will be used to manage file access. - The SFTP protocol will also be used for file distribution to ensure data is encrypted during transmission.
Drop out of a group member	Unforeseen	If a group member drops out of the project, due to the new group size the scope of the project will need to be	High	Low	The project will be split into two substantial sections that do not rely solely on each other.

		changed. This may lead the final product to incur a loss in quality.			
Integration Failure	Operational	The lack of integration among the four main components.	High	Medium	- Design components will be integrated as soon as possible with the addition of system testing after integration. - Additionally components will be designed in a modular fashion with ease of integration in mind.
Security	Legal and Regulatory	Compromising a user's credentials.	High	Low	Usernames and passwords of users will be encrypted in the database.
Ethical Clearance delay	Legal and Regulatory	It is possible that the university has new ethical clearance procedures or the processing of ethical clearance might take some time.	Medium	Low	We will apply for ethical clearance as early as possible in order to avoid any unforeseen delays.

Timeline


2014
May
June
July
August
September
October
November
Outline Project Plan
Completion of Coding
Name
Begin date
End date
Literature Review 2014/05/01 2014/05/15
Project Proposal 2014/05/16 2014/06/16
Project Proposal Due 2014/05/16 2014/05/23
Project Proposal for Zamani 2014/05/16 2014/05/23
Project Proposal Draft 2014/05/16 2014/05/22
Presentation of Project Proposal 2014/05/26 2014/05/28
Presentation Feedback from Staff 2014/05/29 2014/05/29
Revised Proposal Finalized 2014/05/30 2014/06/16
Project Web Presence 2014/06/17 2014/06/17
Outline Project Plan 2014/06/18 2014/06/18
Initial Feasability Study 2014/06/17 2014/07/04
Background/ Theory Chapter 2014/07/07 2014/07/29
First Implementation 2014/07/30 2014/08/29
Sprint 1 2014/07/30 2014/08/07
Sprint 2 2014/08/08 2014/08/15
Sprint 3 2014/08/18 2014/08/29
Final Prototype 2014/09/01 2014/09/18
Sprint 1 2014/09/01 2014/09/09
Sprint 2 2014/09/10 2014/09/18
Completion of Working Prototype 2014/09/19 2014/09/19
Final Implementation 2014/09/19 2014/10/01
Sprint 1 2014/09/19 2014/09/25
Sprint 2 2014/09/26 2014/10/01
Completion of Coding Phase 2014/10/02 2014/10/02
Final Report 2014/10/02 2014/10/29
Final Complete Draft Report Due 2014/10/02 2014/10/22
Final Handin 2014/10/23 2014/10/29
Poster Due 2014/10/30 2014/11/03
Student Web Page Due 2014/10/30 2014/11/05
Completion of Project 2014/11/06 2014/11/06
Project Demonstration 2014/11/06 2014/11/12
Reflection Paper 2014/11/04 2014/11/07
Final Project Presentation 2014/11/10 2014/11/14

Resources Required

Equipment

The project requires the use of computers with access to Zamani's Web server. Additionally a large data storage system, Zamani dataset and metadata is required. All of the equipment required is available and provided by the Geomatics and Computer Science Department.

People

Hussein Suleman is the project supervisor. The development team consists of Michael Ferguson and Jay Benson. The Zamani team consists of Roshan Burtha, Ralph Schroder, Stephen Wessels and Heinz Rüther.

Deliverables

Due Date	Deliverable
15 May 2014	Literature Review
26 May 2014	Project Proposal
15-30 June 2014	Initial Feasibility Study
29 October 2014	Final Project Report
3 November 2014	Project Poster
5 November 2014	Student Web Page
9 November 2014	Reflection Paper

Milestones

Due Date	Milestone
18 June 2014	Outline Project Plan
19 September 2014	Completion of Working Prototype
2 October 2014	Completion of Coding Phase
6 November 2014	Completion of Project

8. References

[1] "Zamani Project". [Online]. Available: http://www.zamaniproject.org. [Accessed: 22-May-2014].

[2] "Zamani Project - Data Types". [Online]. Available: http://www.zamaniproject.org/index.php/data.html. [Accessed: 22-May-2014].

[3] "Zamani Project - Project". [Online]. Available: http://www.zamaniproject.org/index.php/project.html. [Accessed: 22-May-2014].

[4] "Aluka". [Online]. Available: http://www.aluka.org/. [Accessed: 22-May-2014].

[5] Rüther, H., Chazan, M., Schroeder, R., Neeser, R., Held, C., Walker, S. J., Matmon, A. & Horwitz, L. K. Laser scanning for conservation and research of African cultural heritage sites: the case study of Wonderwerk Cave, South Africa. *Journal of Archaeological Science*, 36(9), 2009, 1847-1856.

[6] Rapier, C., & Bennett, B. High speed bulk data transfer using the SSH protocol. In *Proceedings of the 15th ACM Mardi Gras conference*, 2008.

[7] Ilva, J., & Keskitalo, E. P. One process to rule them all? The role of a repository platform in the management of digitized cultural heritage at the National Library of Finland. 2013.

[8] Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley. R., & Walker, J. H. DSpace: An open source dynamic digital repository. *D-Lib Magazine*, 2003.

[9] Lewis, S., Hayes, L., Stangeland, E., Shepherd, K., Jones, R., & Roos, M. DSpace Under the Hood: How DSpace works. In *The 5th International Conference on Open Repositories, 2010*.

[10] Phillips, S., Green, C., Maslov, A., Mikeal, A., & Leggett, J. A New Face for DSpace. *D-Lib Magazine*, 2007.

[11] "DSpace - Discovery". [Online]. Available: https://wiki.duraspace.org/display/DSDOC4x/Discovery. [Accessed: 14-June-2014].

[12] "DSpace - Importing and Exporting". [Online]. Available: https://wiki.duraspace.org/display/DSDOC4x/Importing+and+Exporting+Items+via+Simple+Archive+Format. [Accessed: 14-June-2014].

[13] "DSpace - Submission User Interface". [Online]. Available: https://wiki.duraspace.org/display/DSDOC4x/Submission+User+Interface#SubmissionUserInterface-DefaultSubmissionProcess. [Accessed: 14-June-2014].

[14] "DSpace - Configurable Workflow". [Online]. Available: https://wiki.duraspace.org/display/DSDOC4x/Configurable+Workflow. [Accessed: 14-June-2014].

[15] Matty, S. K. Comparative study of Oracle spatial and postgres spatial. 2012.