

Operational Concept Description (OCD)

Data Mining of Digital Library Usage Data

Team 07

Clients

Jewel Ward

Team Members

Project Manager: Maxim Krivokon

Developer: Bo Lee

Developer: Genesan Kim

Developer: Vu Nguyen

IV&Ver: Shing-Cheung Chan

IV&Ver: Marie Chi

IV&Ver: Kristine Guevara

Version History

“Early Section” releases: 1.x

“LCO Core” releases: 2.x

“LCO Draft on the Web” releases: 3.x

“LCO after ARB” releases: 4.x

“LCA Draft on the Web” releases: 5.x

“RLCA draft” releases: 6.x

“RLCA after ARB” releases: 7.x

Operational Concept Description (OCD) Version no 7.0

Date	Author	Version	Changes made
9/20/04	Shing-Cheung Chan	1.0	<ul style="list-style-type: none"> • Early Section
9/22/04	Shing-Cheung Chan	1.1	<ul style="list-style-type: none"> • Reviewed Early Section
9/28/04	Shing-Cheung Chan	2.0	<ul style="list-style-type: none"> • Full OCD Reports • Modify Results-Chain Diagram
10/8/04	Shing-Cheung Chan	3.0	<ul style="list-style-type: none"> • Reviewed Full OCD Reports • Modify Results-Chain Diagram
10/9/04	Shing-Cheung Chan	3.1	<ul style="list-style-type: none"> • Reviewed Full OCD Reports • Modify Results-Chain Diagram
10/10/04	Shing-Cheung Chan	3.2	<ul style="list-style-type: none"> • Modified Full OCD Reports • Modify Results-Chain Diagram • Modify UML Diagrams
10/16/04	Shing-Cheung Chan	4.0	<ul style="list-style-type: none"> • Review LCO-OCD after ARB • Modify Results-Chain Diagram suggested by the professor • Modify UML Diagrams
10/24/04	Shing-Cheung Chan Maxim Krivokon	4.1	<ul style="list-style-type: none"> • Review LCO-OCD after ARB • Modify Results-Chain Diagram • Modify UML Diagrams • Modify Organization Description suggested by the client
11/11/04	Shing-Cheung Chan Maxim Krivokon	5.0	<ul style="list-style-type: none"> • Review LCO-OCD after IV&V • Modify Results-Chain Diagram • Modify UML Diagrams
11/21/04	Shing-Cheung Chan Maxim Krivokon	5.1	<ul style="list-style-type: none"> • Full LCA-OCD Reports • Modify Results-Chain Diagram • Modify UML Diagrams • Modify Organization and Project Goals Suggested by TAs
11/28/04	Shing-Cheung Chan	5.2	<ul style="list-style-type: none"> • Review LCA-OCD • Modify UML Diagrams
12/03/04	Shing-Cheung Chan	5.3	<ul style="list-style-type: none"> • Review LCA-OCD after ARB • Modify Results-Chain Diagram • Modify UML Diagrams • Modify Glossary List
1/24/05	Bo Lee	6.0	<ul style="list-style-type: none"> • [RLCA REVISION] • Table 14 inserted (Table of tables not updated)
1/30/05	Bo Lee	6.1	<ul style="list-style-type: none"> • [RLCA REVISION] • 4.4 Levels of Services (L.O.S.) Goals revised to be consistent with SSRD 5. LOS Requirements.

Date	Author	Version	Changes made
2/5/05	Bo Lee	6.1.1	<ul style="list-style-type: none"> • [RLCA REVISION] • Added Table 2 for cs577b team. • Inserted team website in the reference. • Updated Table 4 for cs577b team. • Modified 3.3.3 Processes description according to grade comments: "you are saying that they must work together to observe the log data?" on the diagram. (the diagram itself is not updated yet) • New use case for current organization process: evaluate log data linked with project manager and program manager. • Table 9 process-02 added.
2/6/05	Bo Lee	6.2	<ul style="list-style-type: none"> • [RLCA REVISION] • Modified Figure 2,6,7. • Added Figure 8.
2/9/05	Bo Lee	6.3	<ul style="list-style-type: none"> • [RLCA REVISION] • Updated 2.4 PC-2 • Revised 4.4 • Updated Reference
2/22/05	Bo Lee	7.0	<ul style="list-style-type: none"> • [RLCA after ARB] • Revised 2.4 PC-2 • (IV&V) Revised Figure 6. Project manager -> project/program manager (the same person) • (IV&V) Inserted 4.2.5 Limited Resources II heading • Changed the sample log data to new one.

Table of Contents

OPERATIONAL CONCEPT DESCRIPTION (OCD)	I
VERSION HISTORY	III
TABLE OF CONTENTS	VI
TABLE OF TABLES	VIII
TABLE OF FIGURES	IX
1. Introduction	10
1.1 Purpose of the OCD Document	10
1.2 References	11
1.3 Change Summary.....	13
2. Shared Vision	15
2.1 System Capability Description	15
2.2 Key Stakeholders	17
2.3 System Boundary and Environment	18
2.4 Major Project Constraints	19
2.5 Top-Level Business Case.....	19
2.6 Inception Phase Plan and Required Resources	19
2.7 Initial Spiral Objectives, Constraints, Alternatives, and Risks	19
3. Domain/Organization Description	20
3.1 Organization Background	20
3.2 Organization Goals	20
3.3 Current Organization Environment.....	22
4. Proposed System.....	31
4.1 Statement of Purpose	31

4.2	Project Goals and Constraints	31
4.3	System Capabilities	35
4.4	Levels of Service (L.O.S.) Goals	37
4.5	Changes in the Organization Environment Due to Proposed System	38
4.6	Effect on Organizations' Support Operation	46
5.	Prototyping	50
5.1	Objectives	50
5.2	Approach	51
5.3	Initial Results	52
5.4	Conclusions	56
6.	Glossary for Domain Description	57
7.	Appendices	60

Table of Tables

<i>Table 1: Operational Stakeholders Descriptions (CS577a- Fall 2004)</i>	10
<i>Table 2: Operational Stakeholders Descriptions (CS577b – Spring 2005)</i>	11
<i>Table 3: Change Summary</i>	13
<i>Table 4: Key Stakeholder Identifications</i>	17
<i>Table 5: Organization Goal #1</i>	21
<i>Table 6: Organization Goal #2</i>	21
<i>Table 7: Organization Goal #3</i>	21
<i>Table 8: Business Use-Case Description #1</i>	28
<i>Table 9: Business Use-Case Description #2</i>	30
<i>Table 10: Shortcomings</i>	31
<i>Table 11: Project Goals and Constraints #1</i>	32
<i>Table 12: Project Goals and Constraints #2</i>	32
<i>Table 13: Project Goals and Constraints #3</i>	33
<i>Table 14: Project Goals and Constraints #4</i>	33
<i>Table 15: Project Goals and Constraints #5</i>	34
<i>Table 16: Project Goals and Constraints #6</i>	34
<i>Table 17: System Capability #1</i>	36
<i>Table 18: System Capability #2</i>	36
<i>Table 19: System Capability #3</i>	36
<i>Table 20: Level of Service Specification #1</i>	37
<i>Table 21: Level of Service Specification #2</i>	37
<i>Table 22: Level of Service Specification #3</i>	38
<i>Table 23: Business Use-Case Description (New Process) #1</i>	43

Table of Figures

<i>Figure 1: Benefits Realization Approach Results Chain</i>	16
<i>Figure 2: System Boundary and Environment</i>	18
<i>Figure 3: Business-Structure Diagram of Current Organization Structure</i>	22
<i>Figure 4: Business-Collaboration Diagram of Current Organization Structure</i>	23
<i>Figure 5: Business-Artifacts Diagram of Current Organization Artifacts</i>	24
<i>Figure 6 Business Use-Case Diagram of Current Organization Process</i>	26
<i>Figure 7: Business Activity Diagram of Current Organization Process #1</i>	27
<i>Figure 8: Business Activity Diagram of Current Organization Process #2</i>	29
<i>Figure 9: Use-Case Model of System Capability</i>	35
<i>Figure 10:: Business-Structure and Collaboration Diagram of Future Organization Structure</i>	39
<i>Figure 11: Business-Artifacts Diagram of Future Organization Artifacts</i>	41
<i>Figure 12: Business Activity Diagram of Future Organization Process #1</i>	44
<i>Figure 13: Organization Chart #1</i>	47
<i>Figure 14: Organization Chart #2</i>	48
<i>Figure 15: Organization Chart #3</i>	49
<i>Figure 16: Side panel displaying item's meta-data</i>	53
<i>Figure 17: Original unclustered graph</i>	54
<i>Figure 18: Clustered graph</i>	55

1. Introduction

1.1 Purpose of the OCD Document

The Operational Concept Description (OCD) for Data Mining of Digital Library Usage Data describes to the stakeholders how the new USC Digital Library Log Analysis System will function within their environment. This document is the Rebaselined Life Cycle Architecture (RLCA) version of OCD.

The OCD document consists of the scope of the proposed system. Some specific parts are described under different topic in this document. It not only covers the description of the new system, but also includes the primary benefits for client, and agreement on the high level structure of the project of the new system.

The OCD will be served as reference and guidance for all stakeholders as to how the system will operate within its environment. Stakeholders can then make further recommendations on this system, and evolve it to a new operational concept from the current one. The development stakeholders will also have a better understanding of the system, and will make development decisions consistent with the operational objectives and constraints.

The following are the names, organizations, titles, and roles of all operational stakeholders:

Table 1: Operational Stakeholders Descriptions (CS577a- Fall 2004)

Name	Organization	Title	Role
Jewel Ward	USC Library	Digital Resources Librarian	Product Customer/User
Johan Bollen	Old Dominion University	Assistant Professor	Product Customer/User
Shing-Cheung Chan	USC	Student	Developer
Fenny Muliawan	USC	Student	Developer
Hsiao-Han Huang	USC	Student	Developer
Hui-Hsien Chi	USC	Student	Developer
Maxim Krivokon	USC	Student	Developer
Pei Li	USC	Student	Developer
Kristine Guevara	USC	Student	Independent Verification & Validation (IV&V)

Table 2: Operational Stakeholders Descriptions (CS577b – Spring 2005)

Name	Organization	Title	Role
Jewel Ward	USC Library	Digital Resources Librarian	Product Customer/User
Johan Bollen	Old Dominion University	Assistant Professor	Product Customer/User
Maxim Krivokon	USC	Student	Developer
Bo Lee	USC	Student	Developer
Vu Nguyen	USC	Student	Developer
Genesan Kim	USC	Student	Developer
Shing-Cheung Chan	USC	Student	Independent Verification & Validation (IV&V)
Marie Chi	USC	Student	Independent Verification & Validation (IV&V)
Kristine Guevara	USC	Student	Independent Verification & Validation (IV&V)

1.2 References

“Data Mining of Digital Library Usage Data” Project Description

http://sunset.usc.edu/classes/cs577a_2004/projects/description/project7.htm

Model-based Architecting and Software Engineering

http://sunset.usc.edu/research/MBASE/mbase_main.html

MBASE Guidelines v 2.4.2:

http://sunset.usc.edu/classes/cs577b_2005/guidelines/MBASE_Guidelines_v2.4.2.pdf

MBASE Version 2.4.1 Templates for OCD (version 1a)

http://sunset.usc.edu/classes/cs577a_2004/guidelines/MBASEtemplates/OCD_Templatev1a.doc

Visio template for the Results Chain

http://sunset.usc.edu/classes/cs577a_2004/guidelines/MBASEtemplates/agile/ResultsChain_BRA_v2.vst

MBASE Electronic Process Guide

<http://cse.usc.edu/research/MBASE/EPG>

Fall 2003 CS 577a Project #8 LCO portion of OCD

http://ebase.usc.edu/eservices/cs577a_2003/team08a/LCO/OCD_LCO_F03a_T08.pdf

USC Information Services Division

<http://www.usc.edu/isd/about/about.html>

USC Information Services Division—Policies Governing the Use of Computing Resources at USC

<http://www.usc.edu/isd/policies/computing/>

Dr. Johan Bollen's Presentation

http://www.cs.odu.edu/~jbollen/presentations/facstaff_02_28_03.pdf

<http://www.cs.odu.edu/~jbollen/presentations/ecdl02.pdf>

<http://www.cs.odu.edu/~jbollen/presentations/aisti04.pdf>

Digital Archive Visualization Engine

<http://graphics.stanford.edu/~munzner/h3/>

WinWin Spiral Model & Groupware Support System

<http://sunset.usc.edu/research/WINWIN/>

H3Viewer library

<http://graphics.stanford.edu/~munzner/h3/>

CS577b team website

<http://seacliff.usc.edu/~team7b/>

1.3 Change Summary

Table 3: Change Summary

Version	Changes Made (LCO Phrase)
1.1	Rewrite Introduction, Benefits Realized and Domain/Organization Description.
2.0	Complete the rest of Operational Concept Descriptions (OCD) document
3.0	Reflected suggested changes from the OCD Quality Report <ul style="list-style-type: none"> • Reword the Results-Chain Diagram
3.1	Reflected suggested changes from the OCD Early Sections <ul style="list-style-type: none"> • Reword the Results-Chain Diagram
3.2	Reflected suggested changes from team meetings <ul style="list-style-type: none"> • Reword the Results-Chain Diagram
4.0	Reflected suggested changes by the IV&V person, client, and ARB. <ul style="list-style-type: none"> • Include “training” in the Results-Chain Diagram • Modify UML Diagrams to include suggestions from the IV&V person
4.1	Reflected suggested changes from team meetings and ARB <ul style="list-style-type: none"> • Reword Results-Chain Diagram contents • Modify UML Diagrams in section 4.5.3 to include suggestions from ARB • Re-word Organization Description suggested by the client.
	Changes Made (LCA Phrase)
5.0	Reflected suggested changes by the IV&V person and client <ul style="list-style-type: none"> • Modify Results-Chain Diagram (reword intermediate outcomes) • Re-draw UML Diagrams in sections 3.3 and 4.5.3 to illustrate current and future systems
5.1	Reflected suggested changes from team meetings for LCA <ul style="list-style-type: none"> • Re-draw Results-Chain Diagram suggested by the TAs (intermediate and final outcomes) • Modify UML diagrams (actor names) • Redo Organization and Project Goals to stay consistent with the new Results-Chain Diagram
5.2	Reflected suggested changes from team meetings for LCA <ul style="list-style-type: none"> • Modify UML diagrams in section 2.3 and 4.3 to reflect the latest system capability descriptions in section 4.3.

5.3	<p>Reflected suggested changes from LCA-ARB</p> <ul style="list-style-type: none"> • Modify Results-Chain Diagram suggested by the professor (contributions) • Modify UML diagrams (actor names) • Modify UML diagrams in section 4.5.3 to exclude “maintainer” • Eliminated section 2.5, 2.6, 2.7 suggested by the professor • Modify Level-of-service Goals suggested by the professor • Modify Glossary List to include all Domain Descriptions
	<p>Changes Made (RLCA Phrase)</p>
6.0	<ul style="list-style-type: none"> • Update student information for CS577b team • Modified 3.3.3 Processes description according to grade comments from LCA package. • Add new Process-02 to 3.3.3 Processes of current organization diagram. • Add new section 3.3.3.2 Evaluate digital archive usage • Add new UML diagram to 3.3.3.2. • Add new process description to 3.3.3.2
6.2	<ul style="list-style-type: none"> • Modified UML diagrams of Figure 2, 6, 7. • Added UML diagram Figure 8.to 3.3.3.2
6.3	<ul style="list-style-type: none"> • Revised Section 4.4

2. Shared Vision

2.1 System Capability Description

The USC Digital Archive staff has requested an archive usage analysis system as a supplement to standard web metrics such as hit counts. Our proposed usage analysis system will retrieve usage data from the digital library archive and will create relationship graphs for digital archive items. These graphs will give librarians a global structural view of the collections and will help them make decision on updating the collections. This system will be different from other systems because it automates creation of references/relations between digital library uses and documents, and can work with images and other items that do not have metadata or textual description. The system will be applying algorithm from researches conducted by Dr. Johan Bollen of the Old Dominion University. Once the new system is in operation, researchers from various institutions can access the system and requests data and analyses for their own researches.

2.1.1 Benefits Realized

The proposed USC Digital Archive Usage Analysis System has several key benefits. Through visualization of archive collection and usage data (such as IP address, user access time, accessed items, and hit counts), digital archive personnel can gain better understandings of the digital archive collection and usage trends. Based on the information, they can make further improvements to the digital archive. Analyses such as data usage, clustered image evaluations, user trend analysis, and short-term image/document impact ranking versus long-term will assist them in making decisions archive collection improvements.

The proposed system can also benefit Dr. Johan Bollen and other researchers who conduct their works on digital libraries and digital archives. Researchers can request data and analysis reports from the USC Digital Library periodically so that they can apply the results to their research projects.

In summary, the USC Digital Archive Usage Analysis System will benefit the USC Digital Library and digital archive researchers by:

- Better understandings of archive collection structure
- Better understandings of usage trends over time
- References for archive improvements.

2.1.2 Results Chain

The results chain of implementing the digital archive usage analysis system is shown below. The new system will conduct analysis on digital archive item characteristics, which will improve overall understandings of digital archive usage trend. The new system will also conduct analysis on relationships between digital archive items, which will improve the overall understandings on archive collection structure. These help archive staff to make relevant suggestions on improving the digital archive collection, which ultimate improve digital archive services.

The new system also introduces new ways of archive analysis procedure. This means archive staff must be re-trained to adapt to the new procedure. The retraining will ease transition difficulties causes by the deployment of the new system. Ultimately, this will lead to efficient usage of the new analysis system.

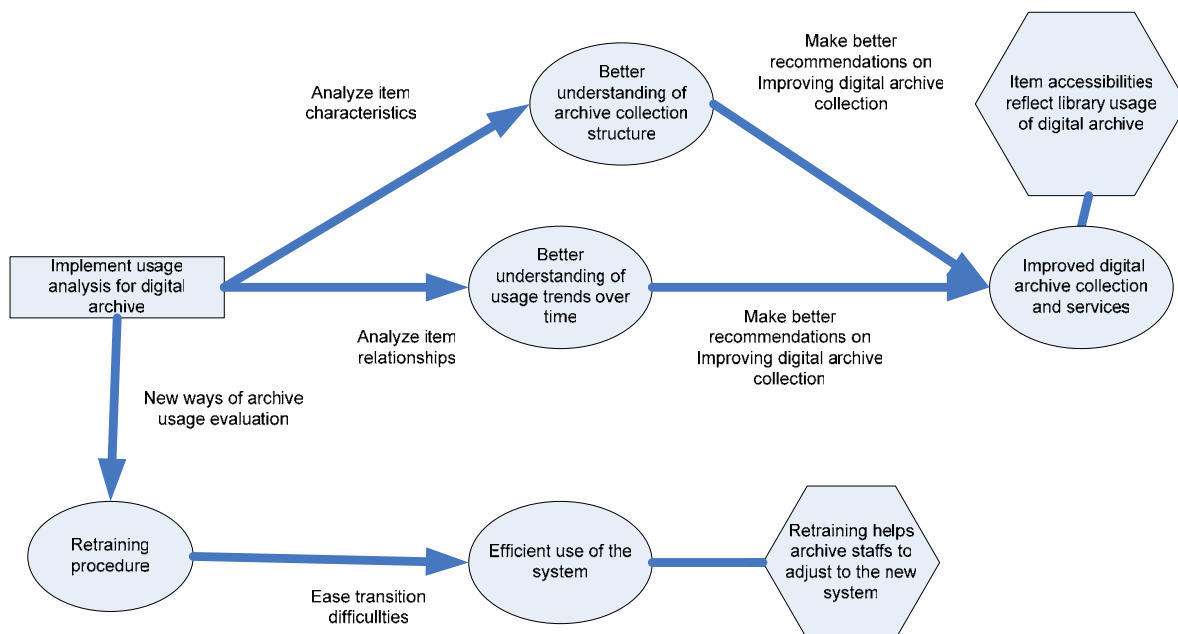


Figure 1: Benefits Realization Approach Results Chain

2.2 Key Stakeholders

The following identifies each stakeholder by their home organization, their authorized representative for project activities, and their relation to the Results Chain.

Table 4: Key Stakeholder Identifications

Stakeholder	Home Organization	Authorized Representative	Relation to the Results Chain
Developer	USC	Maxim Krivokon Bo Lee Vu Nguyen Genesan Kim	Execute the initiative to generate intermediate outcomes
IV&V Person	USC	Kristine Guevara Shing-Cheung Chan Marie Chie	Ensure developers correctly execute the initiative
Maintainer	USC Information Services Division	Jeff Pearson	Ensure necessary data is available for conducting analyses
Program Manager (Customer)	USC Information Services Division	Jewel Ward	Determine the initiative, intermediate outcome, and final outcomes. Experience final outcomes, provide assumptions.
Project Manager	USC Information Services Division	Jewel Ward	Determine the initiative
Researcher	Old Dominion University	Johan Bollen	Provide algorithm for performing data analysis

2.3 System Boundary and Environment

The following diagram for the proposed system includes entities for all the key operational stakeholders in OCD 2.2.

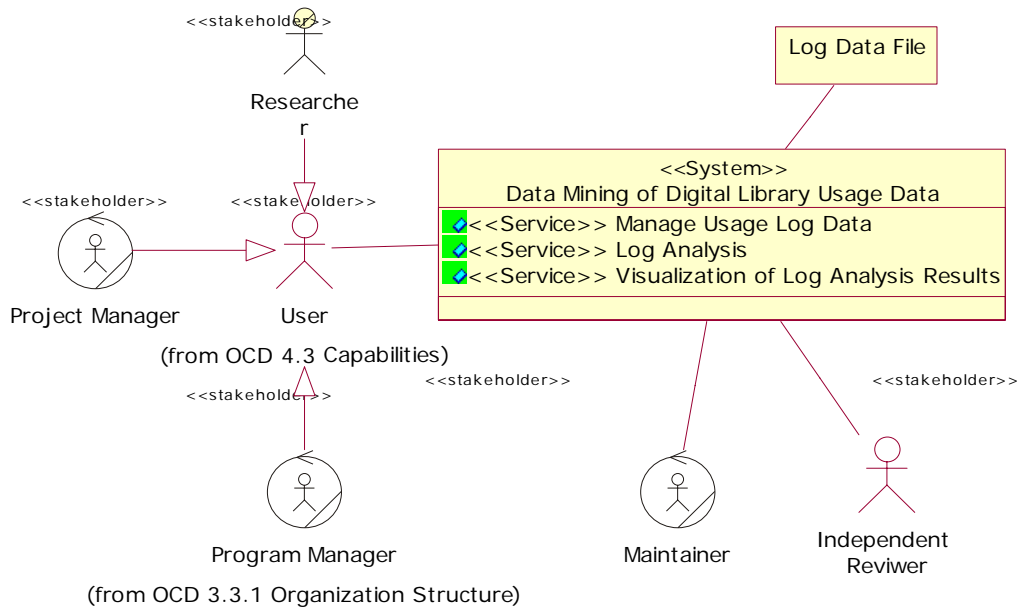


Figure 2: System Boundary and Environment

2.4 Major Project Constraints

There are constraints that are critical to the project's success:

PC-1: Due to course constraints, the project must be completed by the end of spring semester 2005 (roughly 24 weeks).

PC-2: Due to course constraints, the system must be design and implement by six people and two IV&Vers in CS577a and four people and three IV&Vers in CS577b.

PC-3: Due to school constraints, the project developers are not obligated to continue the project in the spring semester of 2005.

PC-4: Due to course constraints, the development of the system will receive no budgets.

PC-5: Due to course constraints, certain project milestones must be reached on specified dates.

2.4.1 Special focus: Further Shared Vision Elements for Large Systems

Special focus is not needed since this is a small development project.

2.5 Top-Level Business Case

This section is not needed since this is a small development project.

2.6 Inception Phase Plan and Required Resources

This section is not needed since this is a small development project.

2.7 Initial Spiral Objectives, Constraints, Alternatives, and Risks

This section is not needed since this is a small development project.

3. Domain/Organization Description

The USC Digital Archive Usage Analysis System is being built to enhance the USC Digital Archive service and to create a user trend and digital item relationships analysis system. Currently, there exist, but not for the USC digital archive, similar usage analysis systems that have been built as part of the recommender services, such as the recommender services for the Los Alamos National Laboratory (LANL) Research Library (RL), the Open Video Project, and the NASA Technical Reports Archive.

However, currently there is no such usage analysis systems built specifically for the USC digital archive. The proposed usage analysis system will assist USC digital archive staff and personnel in evaluating image/document usage and impacts by providing a mechanism with which to cluster related images and analyzing archive user trends. The archive personnel can gain better understandings of digital item characteristics, relationships, and user trends over periods of time. These analyses will contribute to the USC Digital Library's continual efforts to modify archive collections in order to satisfy user needs and improve library services. Currently, archive personnel can only observe raw log data located in the digital archive database and make decisions on archive improvements through regularly-held archive staff meetings. Thus, a brand-new computerized analysis system is desirable which will utilize data and personnel available in the organization.

3.1 Organization Background

The USC Digital Archive is part of the USC Information Services Division (USC-ISD), which is responsible for USC's networking, library services, academic computing, and telecommunications. The Resources and Services Group of the USC-ISD will manage the USC Digital Archive Usage Analysis System and its data collection, while the development of the system will be taken care of by USC graduate students. Researchers from all institutions can also request the information and analyses provided from the USC-ISD and apply to their on-going research. Dr. Johan Bollen of the Old Dominion University is another sponsor of this project. He leads a research group which is conducting researches in data mining of information. The research group would like to apply their data mining and analysis algorithms to real-world application, and then examine the effectiveness of the algorithms.

3.2 Organization Goals

The broad, high-level objectives of the USC Digital Library in relation to Data Mining of Digital Library Usage Data are summarized below:

Table 5: Organization Goal #1

Goal Identifier:	OG-1
Organization Goal:	Improve digital archive collection
Description:	Improve current digital archive collection by gaining better understandings of archive collection structure and usage trend.
Measurable:	Increase in overall usage hit (OCD 2.1)
Relevant:	Having a better understanding of archive collection and usage trend can help archive personnel to improve digital archive. (OCD 3.1)

Table 6: Organization Goal #2

Goal Identifier:	OG-2
Organization Goal:	Retrain archive personnel to simplify organization procedure
Description:	Retrain digital archive personnel in using the new usage analysis system and how to interpret data relationships visualization, so that they can use the system effectively.
Measurable:	New set of training procedure specifically targeting uses of the new system (OCD 2.1)
Relevant:	A set of training procedure designed for archive personnel can ease their difficulties in transitioning to the new computerized procedure from current manual procedures. (OCD 3.1)

Table 7: Organization Goal #3

Goal Identifier:	OG-3
Organization Goal:	Researchers can request usage data and analysis
Description:	Researchers can request data and analysis performed by the system from the USC Digital Archive if they decide to apply analysis results to their own researches.
Measurable:	Researchers can receive copies of results they desired. (OCD 2.1)
Relevant:	Obtaining relevant data and analyses to researches so that researchers can observe the information and draw conclusions. (OCD 3.1)

3.3 Current Organization Environment

The following describes the current organization environment of the USC Information Services Division (USC-ISD). These include the structure, artifacts, processes, rules, and shortcomings of USC-ISD.

3.3.1 Structure

The following describes the current workers of the organization and the outside actors that interact with the organization.

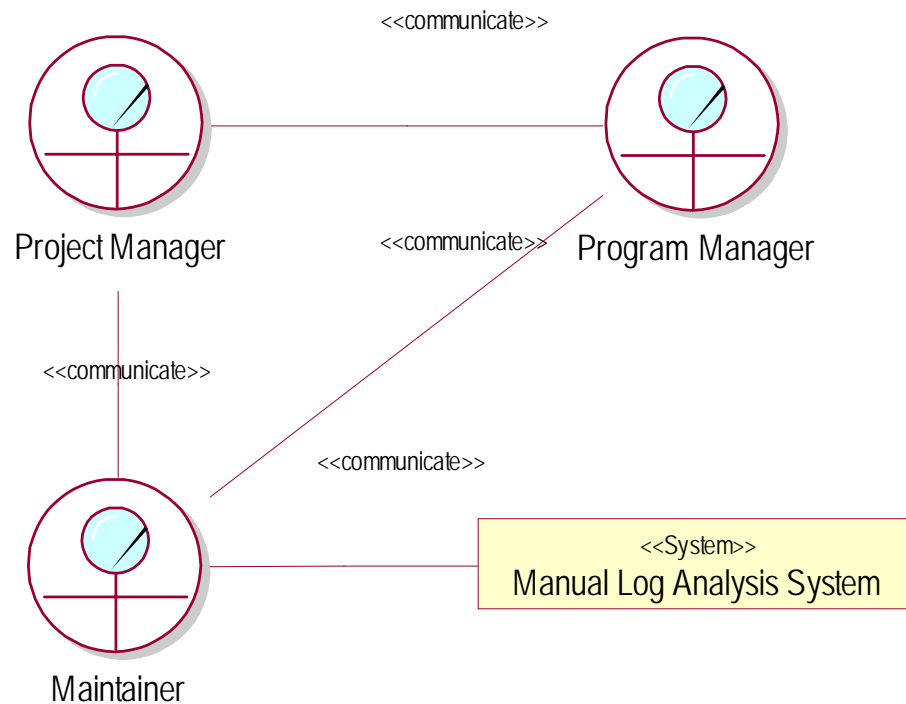


Figure 3: Business-Structure Diagram of Current Organization Structure

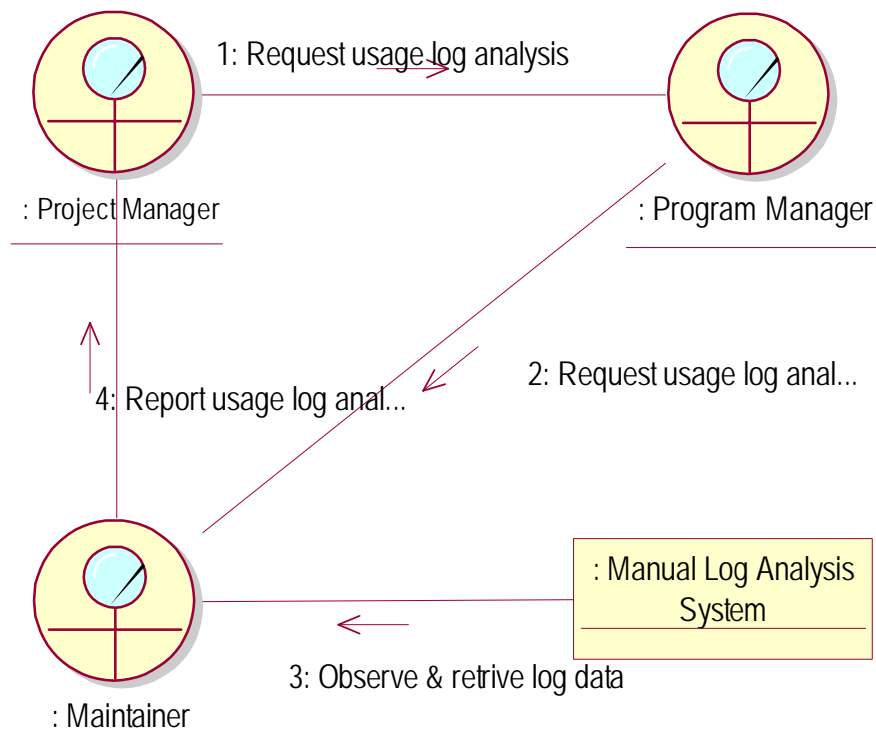


Figure 4: Business-Collaboration Diagram of Current Organization Structure

3.3.1.1 Project Manager

Project managers work in the USC-ISD, and they are responsible for managing projects and systems in USC-ISD. They also manage systems that were acquired from outside resources. Currently, when project managers want to know about the usage of their digital archive system, they make requests to program managers involved in the evaluation processes, and maintainers provide an evaluation report based on their observation on raw log data. The evaluation report contains information such as IP address, user access time, accessed items, and hit counts.

3.3.1.2 Program Manager

Program managers work in the USC-ISD, and they are responsible for overseeing development of the projects and system in USC-ISD. They are also responsible for ensuring successful operations of these projects and systems, and they report their work progresses to project managers. Currently, program managers involving in archive evaluations are responsible assisting maintainers in managing and supporting log data files, which are located In the ISD digital archive log data server. The program managers do not know the information specified in the log data files in great details, thus when project managers request information on archive usage, program managers contact the maintainers to generate an evaluation report based on their observation on raw log data, and the maintainers will send the report directly to the project managers.

3.3.1.3 Maintainer

Maintainers work in the USC-ISD, and they are responsible for managing and observing data in the usage log database. In the current process of archive evaluation, maintainers have exclusive access to the log data server which contains raw log data, and they understand the information specified in the log data files. When project and program managers request information on archive usage, maintainers generate an evaluation report based on their observation on raw log data, such as archive item information, user access times, user IP addresses, and hit counts.

3.3.2 Artifacts

The following describes current artifacts inspected, manipulated, or produced by the organization.



Figure 5: Business-Artifacts Diagram of Current Organization Artifacts

3.3.2.1 Log Data

Log data is a set of files reside in the ISD archive log data server, which includes information such as user's IP, archive items and time of retrieval. The log data files are managed jointly by the USC-ISD program managers and maintainers. However, the regular routine of updating and retrieving log data files in the log data server are done by the maintainers, who are in charged of archive usage evaluations. When digital archive staff request usage information, maintainers will generate reports and distribute to all digital archive staff members during regularly schedule meetings, in which they will discuss about potential improvements that can be made to the digital archive.

3.3.2.2 Usage Analysis Report

The USC-ISD produces digital archive usage reports regularly. Based on these reports, the digital archive personnel can make improvements to the digital archive. The reports usually contain basic information such as archive item information, user access times, user IP addresses, and hit counts, which is provided by maintainers and program managers after observations of data from the usage database. The reports will be distributed to all digital archive staff members during regularly scheduled meetings, in which they will discuss potential improvements that can be made to the digital archive.

3.3.3 Processes

The digital archive staff holds meetings regularly to evaluate data usage and discuss improvements to the digital archive. For the most part, data usage evaluation is performed manually, and evaluations, for example, are based on the number of hits to the digital archive. Currently, this is no computerized usage-analysis system available for the USC-ISD to automate the usage evaluation process. Both program managers and maintainers are responsible for managing log data files, however, maintainers have the sole access to the files in archive log data server. They are knowledgeable enough to interpret the information specified in the log data files, and are solely responsible for observing log data, generating analysis reports, and distributing the reports to project and program managers so that they can evaluate the log data, as well as the rest of digital archive staff.

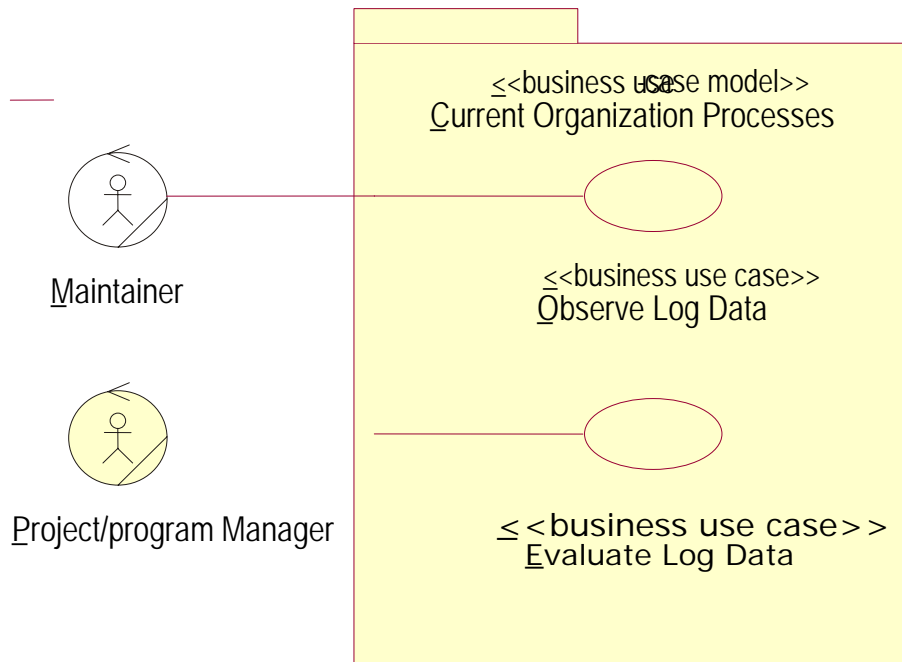


Figure 6 Business Use-Case Diagram of Current Organization Process

3.3.3.1 Observe Log Data

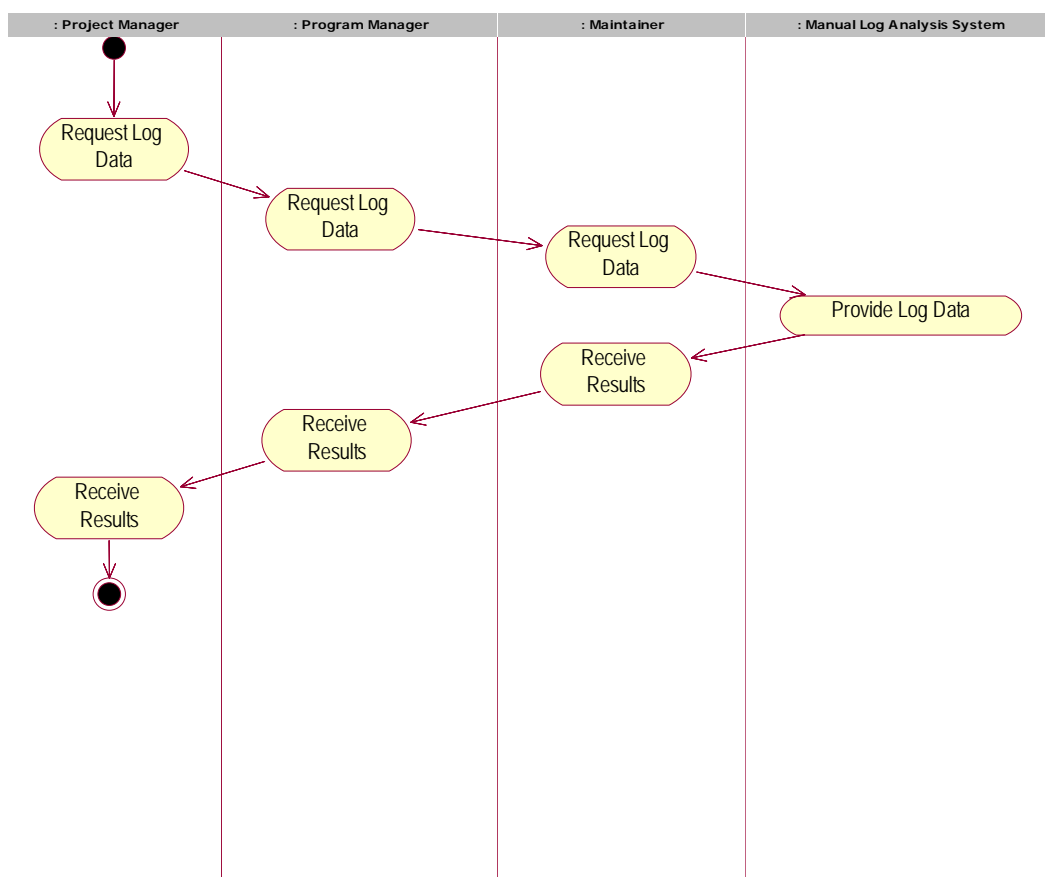


Figure 7: Business Activity Diagram of Current Organization Process #1

Table 8: Business Use-Case Description #1

Identifier	Process-01
Use-Case Name	Observe log data
Purpose	Observe raw log data
Overview	Maintainer observes and reports information related to archive usage to project and program managers, so that they can evaluate digital archive usage and make recommendation on archive improvements.
Organizational Goals	Improve archive collection (OCD 3.2)
Priority	High
Abstract	No
Actors	Maintainer, Program Manager, Project Manager
Pre-conditions	Maintainer can access the database, observe and record information.
Post-conditions	Program and project manager make their evaluations and recommendations on digital archive usage based on that.
Specializes	No
Includes	No
Extends	No
Extension Points	No
Priority	High

3.3.3.2 Evaluate Usage data

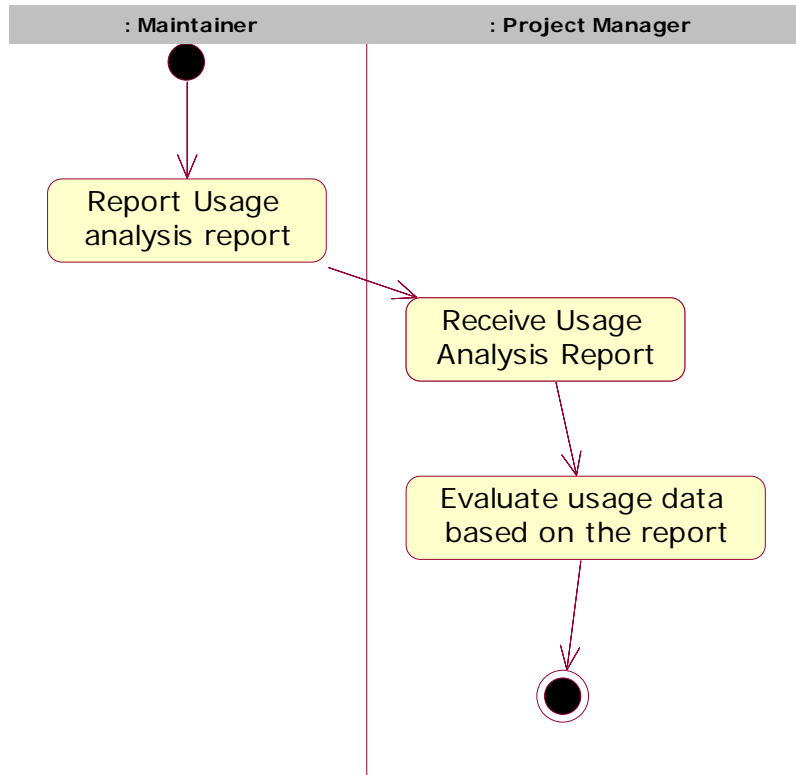


Figure 8: Business Activity Diagram of Current Organization Process #2

Table 9: Business Use-Case Description #2

Identifier	Process-02
Use-Case Name	Evaluate digital archive usage
Purpose	Evaluate digital archive usage derived from log data.
Overview	Project and program managers get the archive usage data from Maintainer and evaluate digital archive usage and make recommendation on archive improvements.
Organizational Goals	Improve archive collection (OCD 3.2)
Priority	High
Abstract	No
Actors	Program Manager, Project Manager, Maintainer
Pre-conditions	Maintainer reports information related to archive usage to Project manager and Program manager..
Post-conditions	Program and project manager make their evaluations and recommendations on digital archive usage based on that.
Specializes	No
Includes	No
Extends	No
Extension Points	No
Priority	High

3.3.4 Rules

R-1: The organization must satisfy copyright law, which prohibits distribution or reproduction, in any digital form, of copyrighted music, video, or other multimedia contents without the express written permission of the material's rightful owner.

R-2: Anyone in the USC-ISD cannot grant access to the digital archive log data server to any outsider without permission.

R-3: Within the USC-ISD, anyone who wants to access the digital archive log data server must notify both program managers and maintainers in charge of the log data server.

R-4: Log analysis reports can only be generated by the maintainers of the log data server. Reports generated by other ISD personnel are invalid.

3.3.5 Shortcomings

The following is a list of shortcoming of the current procedure of usage evaluations:

Table 10: Shortcomings

ID	Shortcomings	Descriptions
SC-1	Not enough information for thorough evaluation	The currently usage analysis only based on elementary information such as IP address, user access times, item being accessed, and hit counts as the basis of evaluation. This does not provide archive staff with meaningful information about archive usage trends, and thus is not very helpful in enhancing library services.
SC-2	Lack of direct access	The current procedure shows that only program managers and maintainers have accesses to the usage database. Project managers and other archive staff have no direct access.

4. Proposed System

This section depicts the proposed system for Data Mining of Digital Library Usage Data and explains what the system is, what it performs and how well it performs.

4.1 Statement of Purpose

The USC archive staff needs an archive usage analysis system which can provide a more detailed analysis than a simple manual usage evaluation procedure. Our proposed system can address current system shortcomings can retrieve necessary data from the digital archive and perform suitable analysis on user trends, impact ranking, and image/document clustering. These analyses will have digital archive staff to gain in-depth knowledge about digital archive usage trends, archive collection structure, and item relationships. These will help digital archive staff to make further improvements to the digital archive collection. Furthermore, the new computerized system gives ISD archive personnel, such as project managers, program managers, and maintainers, direct access to the usage analysis system. Also, researchers can request data and analyses done by the system to assist them for their on-going researches. They can also request the USC Digital Archive for direct accesses to the usage analysis system. In short, the new digital archive usage analysis system will benefit everyone interested and involved in the digital archive community.

4.2 Project Goals and Constraints

The main goals and constraints that are critical to the project's success are described below:

4.2.1 Better Understandings of Archive Collection Structure

Table 11: Project Goals and Constraints #1

Project Goal:	PG-1: Better Understandings of Archive Collection Structure
Description:	The project will develop a system which gives archive staff better understandings of archive collection structure.
Measurable:	The system can provide more relevant statistics than current procedure.
Relevant:	Compatible with improving archive collection goals (OCD 3.2, OG-1) Compatible with not enough information for thorough evaluation shortcoming (OCD 3.3.5, SC-1)
Specific:	Improve current digital archive collection by gaining better understandings of archive collection structure and usage trend. (OCD 3.2, OG-1) The currently usage analysis only based on elementary information such as IP address, user access times, item being accessed, and hit counts as the basis of evaluation. This does not provide archive staff with meaningful information about archive usage trends, and thus is not very helpful in enhancing library services. (OCD 3.3.5, SC-1)

4.2.2 Better Understandings of Archive Usage Trends

Table 12: Project Goals and Constraints #2

Project Goal:	PG-2: Better Understandings of Archive Usage Trends
Description:	The project will develop a system which gives archive staff better understandings of archive usage trends.
Measurable:	The system can provide usage trends data accurately.
Relevant:	Compatible with improving archive collection goals (OCD 3.2, OG-1) Compatible with not enough information for thorough evaluation shortcoming (OCD 3.3.5, SC-1)
Specific:	Improve current digital archive collection by gaining better understandings of archive collection structure and usage trend. (OCD 3.2, OG-1) The currently usage analysis only based on elementary information such as IP address, user access times, item being accessed, and hit counts as the basis of evaluation. This does not provide archive staff with meaningful information about archive usage trends, and thus is not very helpful in enhancing library services. (OCD 3.3.5, SC-1)

4.2.3 Efficient usage analysis procedure

Table 13: Project Goals and Constraints #3

Project Goal:	PG-3: Efficient usage analysis procedure
Description:	Digital archive staff will use the new system efficiently
Measurable:	Digital archive staff can use the system to conduct analysis and product reports.
Relevant:	Compatible with retrain archive personnel to simplify procedure goals. (OCD 3.2, OG-2) Compatible with Lack of direct access shortcoming. (OCD 3.3.5, SC-2)
Specific:	A set of training procedure designed for archive personnel can ease their difficulties in transitioning to the new computerized procedure from current manual procedures. (OCD 3.2, OG-2) The current procedure shows that only program managers and maintainers have accesses to the usage database. Project managers and other archive staff have no direct access. (OCD 3.3.5, SC-2)

4.2.3 Fixed Schedule

Table 14: Project Goals and Constraints #4

Project Constraint:	PC-1: Fixed Schedule
Description:	The project must be completed by the end of spring semester 2005, implemented by six students, and reached project milestones on time.
Measurable:	The project can be delivered on schedule
Relevant:	Compatible with fixed project completion time and fixed number of team developers constraints (OCD 2.4, PC-1, PC-2, PC-3, PC-5)
Specific:	(OCD 2.4, PC-1): Due to course constraints, the project must be completed by the end of spring semester 2005 (roughly 24 weeks). (OCD 2.4, PC-2): Due to course constraints, the system must be design and implement by six people and two IV&Vers in CS577a and five people and three IV&Vers in CS577b. (OCD 2.4, PC-3): Due to school constraints, the project developers are not obligated to continue the project in the spring semester of 2005. (OCD 2.4, PC-5): Due to course constraints, certain project milestones must be reached on specified dates.

4.2.4 Limited Resources I

Table 15: Project Goals and Constraints #5

Project Constraint:	PC-2: Limited Resources
Description:	The development of the system will receive no budgets and outside assistances.
Measurable:	The project can move along without receiving budgets or outside helps.
Relevant:	Compatible with development of system will receive no budgets, with fixed number of developers constraints. (OCD 2.4, PC-2, PC-4)
Specific:	<p>(OCD 2.4, PC-2): PC-2: Due to course constraints, the system must be design and implement by six people and two IV&Vers in CS577a and five people and three IV&Vers in CS577b.</p> <p>(OCD 2.4, PC-4): Due to course constraints, the development of the system will receive no budgets.</p>

4.2.5 Limited Resources II

Table 16: Project Goals and Constraints #6

Project Constraint:	PC-3: Development environment limited to Open source (GNU)
Description:	<p>With the constraint depicted as PC-2. Development environment is limited to Open source. Configuration and compilation of the system source code should be automated using GNU tools</p> <p>(such as GNU Make of version 3.80 and up. Other tools such as autoconf, automake is optional.: THIS PART TOO SPECIFIC FOR OCD)</p>
Measurable:	<p>The development environment is based on GNU utility program.</p> <p>(SSRD: The produced code is compiled by running the corresponding GNU utility program.)</p>
Relevant:	<p>Ensures ease of system installation. Realizes win condition 3.3</p> <p>Compatible with zero budget constraint (OCD 2.4, PC-4)</p>
Specific:	(OCD 2.4, PC-4): Due to course constraints, the development of the system will receive no budgets.

4.3 System Capabilities

The new USC Digital Archive Usage Analysis System will provide a brand-new usage analysis system to all digital archive staff and personnel with capabilities that satisfy the majority who are interested and involved in the digital archive community. The capabilities listed below are broad categories of the analysis system, and these should realize the high-level services described in the System Boundary and Environment (OCD 2.3). Specifically, the new usage analysis system will be an individual stand-alone desktop-based system. Users (project managers, program managers, and researchers) will import log data files manually to the system, which it will perform usage analysis and visualizations. Afterwards, users can save the results (analysis reports) to a client database located in the same machine (the local database will be distributed along with the usage analysis system upon requested). When any digital archive staff or researchers request analysis reports, the users can retrieve relevant information from that local database. When the new usage analysis system is in operation, the currently procedure of log data observation that involve passing messages and collaboration among digital archive staff will be eliminated.

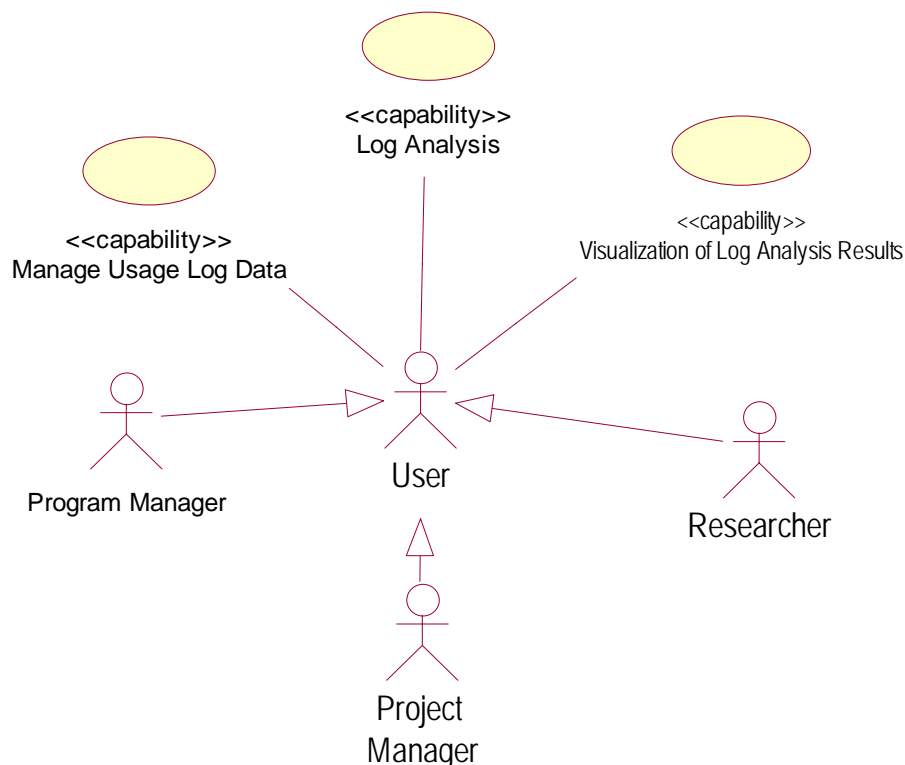


Figure 9: Use-Case Model of System Capability

4.3.1 Manage Usage Log Data

Table 17: System Capability #1

Identifier	C-01
Name	Manage Usage Log Data
Description	Import/export archive usage data provided by the user and system
Importance	Primary
Use In	Usage Log Analysis Report Generation (OCD 4.5.3)
Priority	High. This is the essential component of the archive usage analysis.

4.3.2 Log Analysis

Table 18: System Capability #2

Identifier	C-02
Name	Log Analysis
Description	Use algorithm proposed by Dr. Johan Bollen to analyze digital archive items
Importance	Primary
Use In	Usage Log Analysis Report Generation (OCD 4.5.3)
Priority	High. This is the essential component of the archive usage analysis.

4.3.3 Visualization of Log Analysis Results

Table 19: System Capability #3

Identifier	C-03
Name	Visualization log analysis results
Description	Generate graphical views of item relationships and collection structure
Importance	Primary
Use In	Usage Log Analysis Report Generation (OCD 4.5.3)
Priority	High. This is the essential component of the archive usage analysis.

4.4 Levels of Service (L.O.S.) Goals

The following describes the levels of service goals for the usage analysis system.

Table 20: Level of Service Specification #1

Level of Service:	LS-1: System Dependability
Description:	System should operate without crash or unexpected restart.
Measurable:	Performance statistics Error log.
Relevant:	Usage analysis system should be reliable (OCD 4.3)
Specific:	System regularly performs usage log analysis (OCD 4.3)
Priority:	High

Table 21: Level of Service Specification #2

Level of Service:	LS-2: Usability
Description:	System interface should be easy to use for non-technical users. System should take up minimum resource of host computer for maximum usability of available resources when analyzing and visualizing the usage log data.
Measurable:	Usability test feedback. Number of tasks a user can accomplish per unit time on a host computer. (usability test)
Relevant:	Importing usage data should be easy. (OCD 4.2) Viewing the result of usage analysis should be easy and not take up too much resources of host computer. (OCD 4.3) Efficient usage analysis procedure ensures the efficient use of the resources. (OCD 4.2)
Specific:	Bad interfaces mislead users (OCD 4.3) Visualization and analysis of log data can consume too much resource (OCD 4.3)
Priority:	High

Table 22: Level of Service Specification #3

Level of Service:	LS-3: Performance
Description:	The performance of the system is based on data of current scale, and the data should be organized in the way that is meaningful to both archive staff and researchers.
Measurable:	The system should be able to work with the desired number of nodes for current condition. Archive staff will provide usability feedback Data successfully apply to research publications
Relevant:	Provide meaningful datasets (OCD 4.2)
Specific:	Data must be presented in a meaningful way so that it can be used as a basis for researches (OCD 4.2)
Priority:	Medium

4.5 Changes in the Organization Environment Due to Proposed System

The following sections describe how the new usage analysis system changes the environment of the organization. This includes the structure, artifacts, processes, rules, and how the new system remedies the shortcomings of the organization's current environment and system.

4.5.1 Structure

The following sections describe how the new usage analysis system will change organization's architecture.

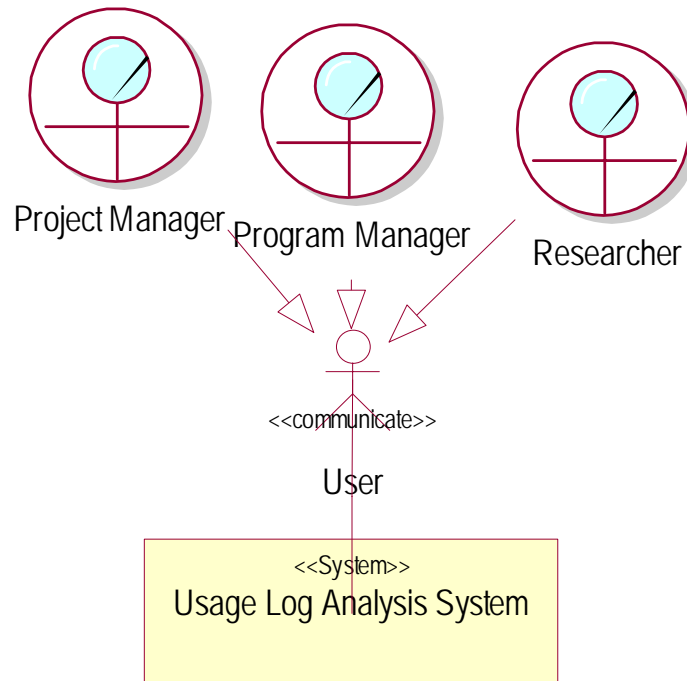


Figure 10:: Business-Structure and Collaboration Diagram of Future Organization Structure

4.5.1.1 New Workers and Outside actors

4.5.1.1.1 *Researcher*

Researchers who conduct researches in digital archives can request analyses and reports on archive usage analysis from the USC-ISD. Researchers will be able to acquire information they seek for from program managers. Researchers can also request direct accesses to the usage analysis system by requesting a copy of the stand-alone system, and they can run the system the same way as the project and program manages. However, they still need to request the USC-ISD for copies of log data files.

4.5.1.2 Changed Workers and Outside actors

4.5.1.2.1 Project Manager

Project managers will have direct access to the usage analysis system (by requesting copies of the system and the local database) so they can view various statistics and relationship diagrams on screen. They will no longer need to request information through program managers. Project managers can gain full access to the system capabilities described, and run the new system individually the same way as program managers.

4.5.1.2.2 Program Manager

Program managers will have direct access to the usage analysis system (by requesting copies of the system and the local database) so they can view various statistics and relationship diagrams on screen. They will no longer receive analysis requests from project manager, and they will no longer need to request information through maintainers. Program managers can gain full access to the system capabilities described, and run the new system individually the same way as project managers.

4.5.1.3 Deleted Workers and Outside actors

4.5.1.3.1 Maintainer

Maintainers can maintain their full access to the usage analysis system (by requesting a copy of the stand-alone system), but they will no longer involve in digital archive usage evaluation process within the organization. When the new system is in operation, maintainers' key role will be to maintain the log data files resided in the digital archive log data server (outside of the new usage analysis system), so that users of the analysis system can request valid copies of log data files at all times.

4.5.2 Artifacts

The following sections describe how the new usage analysis system will change the organization's artifacts.



Figure 11: Business-Artifacts Diagram of Future Organization Artifacts

4.5.2.1 New Artifacts

4.5.2.1.1 Automated Usage Analysis Report

The USC-ISD will automate digital archive usage reports by using the proposed usage analysis system. The system will produce reports based on information in the log data files and manipulate various statistics such as impact ranking and usage trends over time. Digital archive staff can make recommendations for archive improvements based on information gathered in the reports. These analysis reports can be saved to a locally-hosted database for later retrieval.

4.5.2.2 Changed Artifacts

4.5.2.2.1 Log Data

The process of retrieval of log data files will be changed when the new system is in operation. Currently, log data files are retrieved and observed directly by the maintainers. However, in the new usage evaluation process the log data files can be retrieved by any users subscribed to the new usage analysis system.

4.5.2.3 Deleted Artifacts

4.5.2.3.1 *Usage Analysis Report*

The USC-ISD will automate digital library usage reports by using the proposed usage analysis system. Thus, the manually generated usage analysis reports prepared by the maintainers will be eliminated from the evaluation process.

4.5.3 Processes

The following sections describe how the new usage analysis system will change the organization's processes. A new process has been added because of the system.

4.5.3.1 New Processes

The following describes organization processes that have been added because of the new usage analysis systems

4.5.3.1.1 *Usage Log Analysis and Report Generation*

Table 23: Business Use-Case Description (New Process) #1

Identifier	Process-10
Use-Case Name	Usage Log Analysis and Report Generation
Purpose	Generate usage data analysis reports between a specific timeframe
Overview	Report the analyses conducted and displayed by the usage analysis system
Organizational Goals	Improve digital archive collection (OCD 3.2, OG-1) Requests usage data and analyses (OCD 3.2, OG-3)
Priority	High
Abstract	No
Actors	Users (project managers, program managers, and researchers)
Pre-conditions	Users want to evaluate digital archive usage; log data files are ready to import.
Post-conditions	Usage analysis is completed, report is generated for users, and they are opted to save the reports.
Specializes	No
Includes	No
Extends	No
Extension Points	No
Priority	High

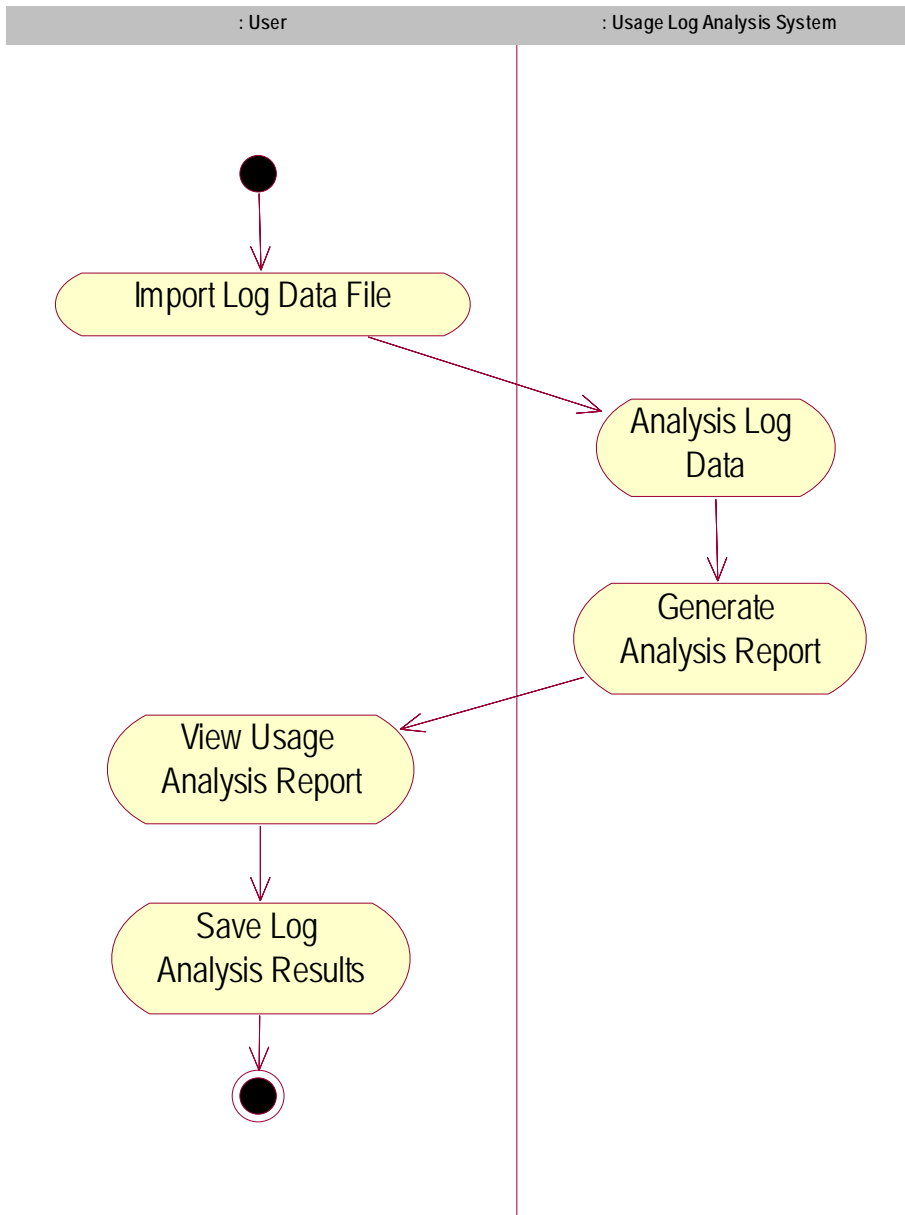


Figure 12: Business Activity Diagram of Future Organization Process #1

4.5.3.2 Changed Processes

There are no changed processes.

4.5.3.3 Deleted Processes

The following describes organization processes that have been deleted because of the new usage analysis systems.

4.5.3.3.1 *Observe Log Data*

The new process “Usage Log Analysis Report Generation” automates the evaluation of digital archive and thus the current process of manual log data observation is unnecessary.

4.5.4 Rules

The following sections discuss any changes to the policies or constraints that must be satisfied by the organization.

4.5.4.1 New Rules

The following new rules will be enforced:

NR-1: Everybody have access to the usage system.

NR-2: Maintainers will not take part in usage evaluation process.

4.5.4.2 Changed Rules

The following rules will be changed:

R-3: Currently within the USC-ISD, anyone who wants to access the digital archive log data server must notify both program managers and maintainers in charge of the log data server. In the future, anyone can ask for a copy of the new stand-alone usage analysis system from the USC-ISD.

R-4: Currently, log analysis reports can only be generated by the maintainers of the log data server. Reports generated by other ISD personnel are invalid. In the future, reports generated by those who subscribe to the system and import the log data files from the archive log data server will be considered as valid.

4.5.4.3 Deleted Rules

No rules will be deleted.

4.5.5 How New System Cures Current Shortcomings

Successful development and installation of the proposed system will address the shortcomings by granting everyone full access to the usage analysis system so that they can run analysis on digital archive usage. The new system will eliminate excessive communications among digital archive staff. **(SC-1)**

Successful development and installation of the proposed system will also address the shortcomings by providing a justification as to which digital documents are more relevant than others and which documents have been accessed by most digital archive users. This helps digital archive staff to accurately evaluate digital archive usage and enhance services. **(SC-2)**

4.6 Effect on Organizations' Support Operation

4.6.1 Operational Stakeholders

The figure below shows the responsibility relations between the various organizations involved in the software life cycle process, and identifies the key responsible personnel within each organization.

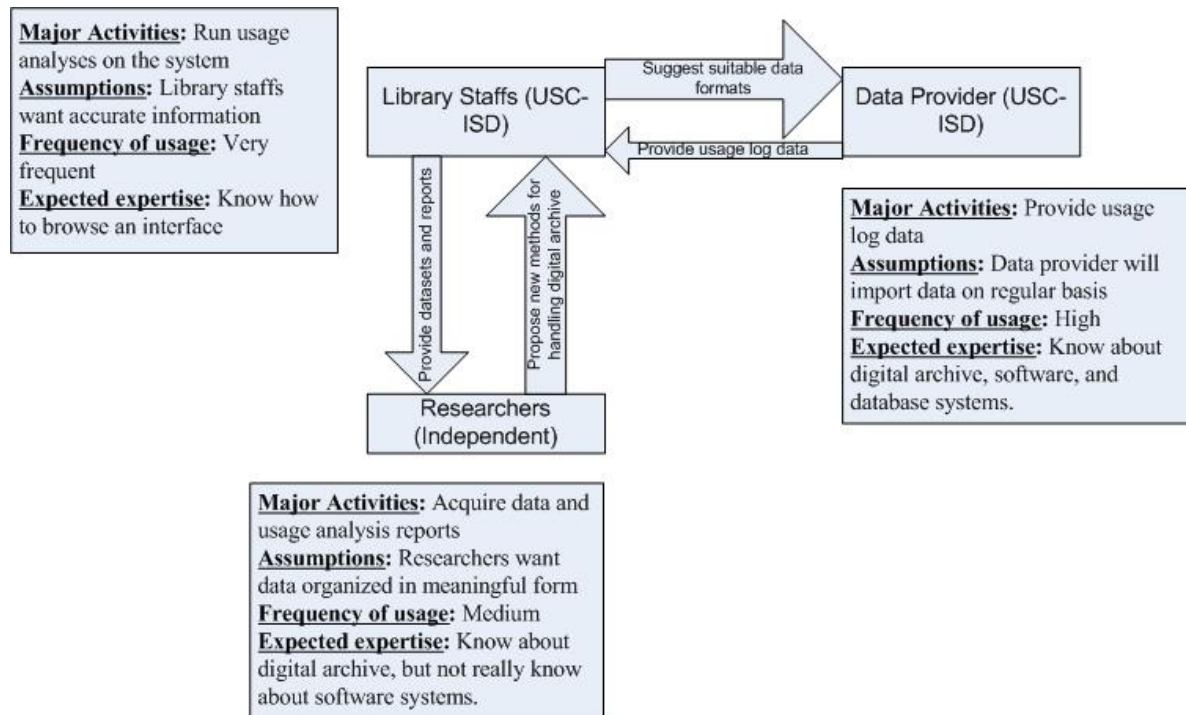


Figure 13: Organization Chart #1

4.6.2 Organizational Relationships

The USC-ISD is responsible for providing computing support and electronic services to the USC community. Specifically, the Information Data Management Department (IDM) under the USC-ISD is responsible for managing projects and systems, which include the digital library and the digital archive. Within IDM, there are project sponsors, project managers, program managers, directors, and implementation teams. They are responsible for funding, managing, overseeing, directing, and implementing projects, respectively. The USC-ISD is committed to bringing the USC community high quality technological services, and IDM realizes such commitment by building and maintaining computer systems that satisfy the needs of the USC community. An example of this is to provide and enhance digital library services.

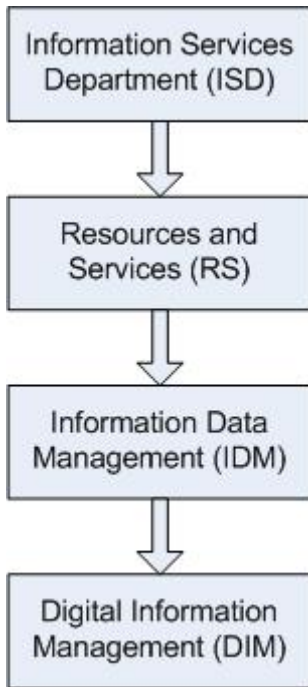


Figure 14: Organization Chart #2

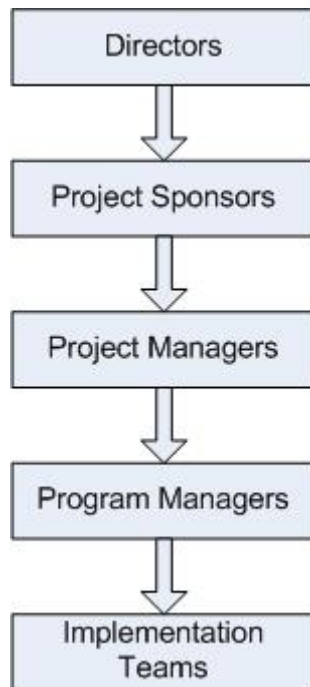


Figure 15: Organization Chart #3

4.6.3 Transition & Evolution

Potential impacts of the new operational concept on operational personnel, procedures, performances and management functions due to the parallel operation of the new and existing systems are listed below:

- Digital archive needs to have archive personnel to regularly retrieve usage analyses results
- Digital archive needs to have archive personnel to handle usage analyses system maintenance
- Digital archive improvements will be made with consideration given to the results of the usage analysis system.
- Digital collections will be added to or enhanced based on the usage patterns of each collection.
- Existing usage observation method will continue until the new system is in-place.

4.6.4 Operation Support

Once the system is in operation, USC Digital Archive staff will have a powerful usage analysis tool to assist them in making improvements to the digital library archive. They will use the results of the usage analyses performed by library faculty and/or staff to make modifications to the digital archive, and thus enhance digital archive collection and services. Manual observation/evaluation of data will be eliminated.

5. Prototyping

5.1 Objectives

5.1.1 Visualization prototype

The purpose of our prototype is to demonstrate visualization of relationships and associated information produced by analysis of Digital Archive usage logs. During Win-win negotiations development team and client have reached common understanding of analysis algorithm to be used which is based on the research of Johan Bollen. Thus the team decided to develop a visualization rather than functional prototype because we consider user interface to be area of highest risk in the project. Choice of visualization framework affects other system components in the following ways:

- visualization framework determines system architecture. (e.g. is system is chosen to have static visualization – that is only pictures then it makes sense to make it web-based, however if dynamic visualization interface is required then it is more suitable to design system as stand-alone program)
- since visualization module will be querying analysis results, it determines in what format those results should be stored
- requirements to the information that should be displayed determine what kind of analysis should be performed and what should be the contents of analysis results

Also this prototype is meant to address the IKIWISI syndrome that our project experienced in defining user interface requirements. The type of system that is being proposed in this project is relatively new and has not been used by the client before. Thus, during the negotiations it was difficult to arrive at detailed description of how the system should look like, what kind of options user should be given, and how information should be visualized. This prototype is meant resolve this syndrome and propose interface design elements that might be used in the system. Evaluation and testing of these interface examples should produce a vision of the future system that will satisfy the client and will be deemed feasible by the development team.

Besides user interface requirements uncertainties the prototype addresses the following visualization issues that stem from specifics of the source data:

- To many graph nodes: current collection contains around 300000 items and all of them will have to be represented as graph nodes
- To many edges: item relationships are scattered and numerous due to the nature of the data the relationships are produced from (usage logs are relatively irregular)
- How to visualize high level collection structure based on item relationships. Showing all item relationships will clutter the view a lot and will not help to understand collection's structure.
- How to combine graph view with detailed information about nodes of interest.

5.2 Approach

5.2.1 Scope and Extent

The prototype is partially functional: some capabilities are demonstrated in action using open source component, and some are presented as mockup pictures. The following desired capabilities have been prototyped:

- 3d hyperbolic view of relationship graph
- Multi-level graph clustering, and tree generation
- Side-bar for item's meta-data

User interface demonstrated by this prototype has been designed to tailor client's needs, and results of the development were reviewed by the client. Thus this prototype manifests an agreement between the development team and the client on how the system interface should look like. The visualization platform that has been chosen for this prototype might become the development platform for the system after additional evaluation and testing.

5.2.2 Participants

Development of the prototype was based on the results of Win-Win negotiations which were conducted by the client and the team. During the development intermediate results were provided to the client for evaluation and team meetings were set up to discuss further development plans. Two team members were responsible for the prototype deliverables: Fenny Muliawan and Maxim Krivokon.

5.2.3 Tools

As a result of research in the information visualization area H3viewer was chosen as a graph visualization platform. It is an open-source library that provides graph visualization in 3d hyperbolic space allowing to display large graphs in uncluttered manner. This prototype uses a sample viewer program that is developed based on H3viewer library and is available from H3Viewer website. Standard office graphical tools, such like MS Paint, MS Excel have been used to develop a mock up screen of a side panel that provides detailed view of item information.

5.2.4 Revision History

Date	Author	Version	Changes made
09/26/2004	Fenny Muliawan	1.0	Preliminary Draft
09/30/2004	Fenny Muliawan	1.1	Changes in requirement (Web-based is no longer required)
10/23/2004	Maxim Krivokon	1.2	Introduce new screenshots. Reorganize some sections
11/21/2004	Maxim Krivokon	2.0	Add new risks

5.3 Initial Results

H3viewer library provides the following visualization features:

- detailed, magnified view at point of interest
- shrinking of nodes and edges at periphery
- shows only spanning tree edges

These features address issues A and B. Hyperbolic view is meant to unclutter the view by devoting most of the visual space only to the small part of the graph that interests the user, and presenting the rest of the graph in distorted minimized perspective. Even if the graph has many nodes this visualization capability allows user to concentrate attention only the small subset of nodes. The issue B is worked around by not showing all edges but only edges that produce spanning tree of the graph. In that way connectivity of the graph is preserved and the view is uncluttered at the same time.

However according to stated objectives H3viewer has the following issues:

- only name of a node can be displayed
- spanning tree edges (Depth First Search or Breadth First Search) do not reflect collection's structure

These two mirror issues C and D identified in section 5.1. To overcome these issues we propose the following extensions to the existing H3viewer visualization tool:

- *Display meta-data associated with each node*
User interface of the tool will have to be modified to provide additional space for displaying this information. We propose to create a side-panel to the right of graph browsing window that would display meta-data of a selected item. (See Figure 1)
- *Provide view that conveys clustering of nodes based on their relatedness.*
We propose to generate a hierarchy tree based on multilevel clustering of original relationship graph in the following way: (See Figures 2, 3)
For a given level x (where level = sub graph)
 - Find the most important item in the level (item that is related to most items and is being related to by most items)
 - Make it the child of this level's parent
 - Cluster all the remaining nodes in this level
 - Make the resulting clusters children of the most important item

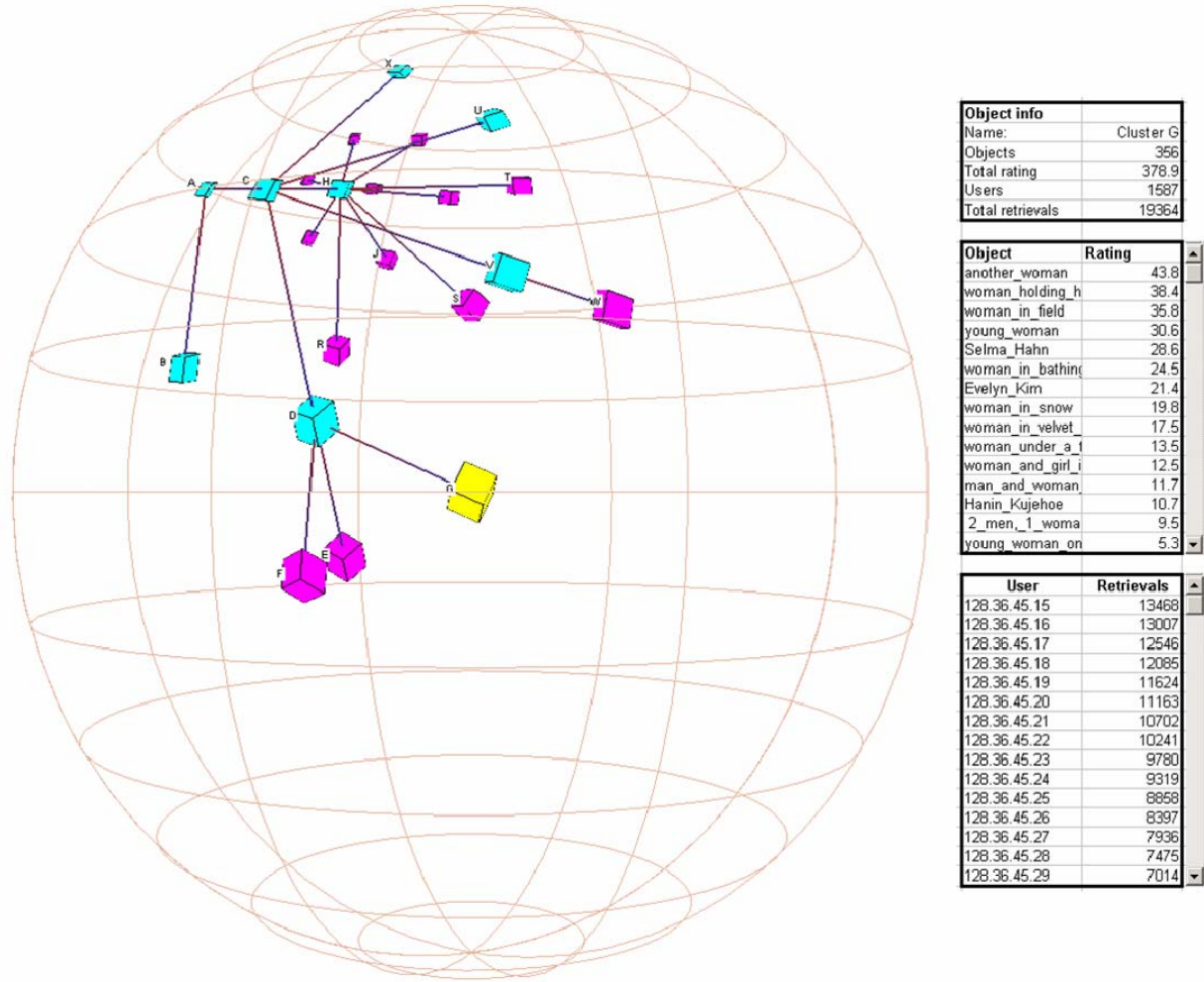


Figure 16: Side panel displaying item's meta-data

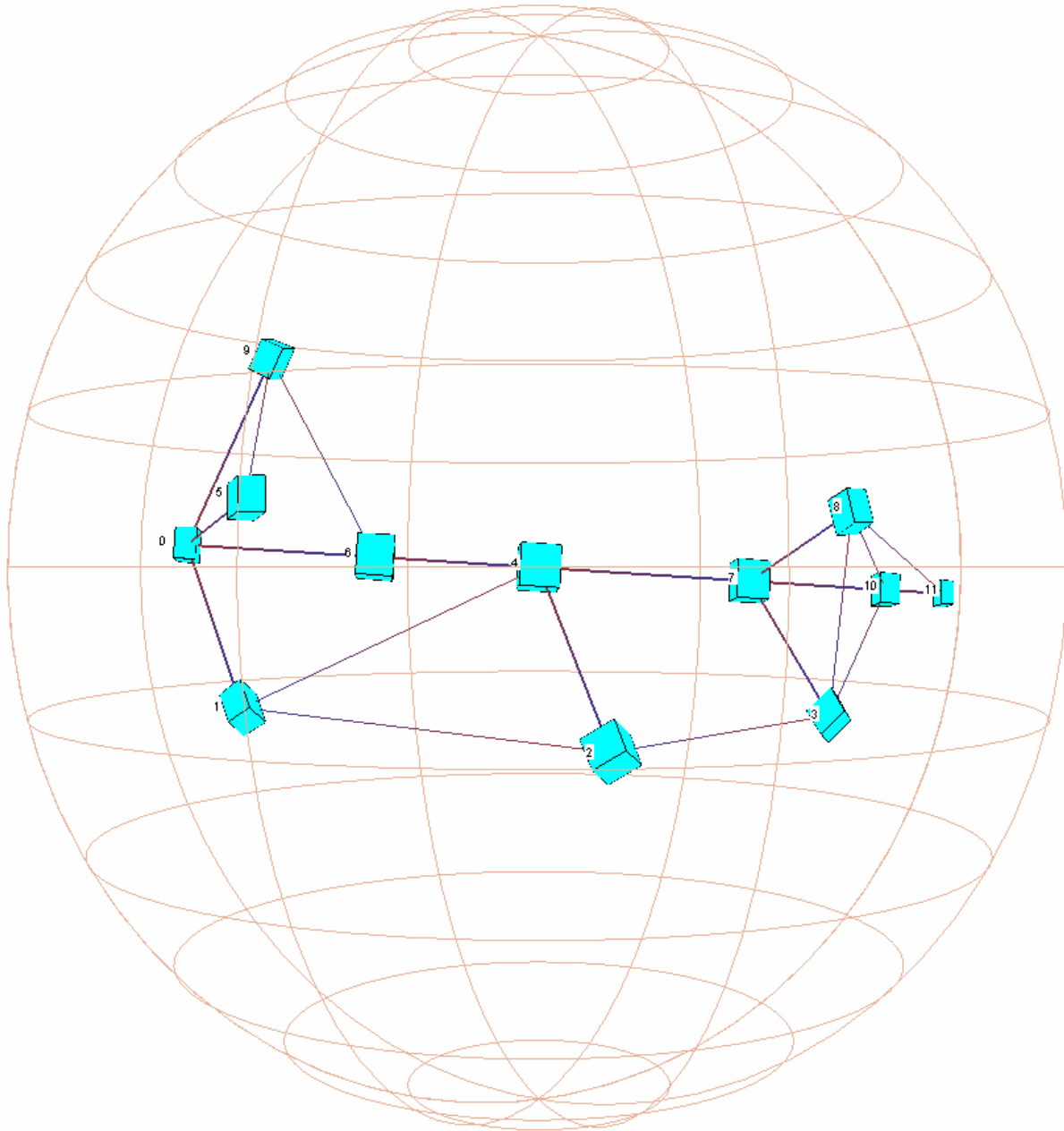


Figure 17: Original unclustered graph

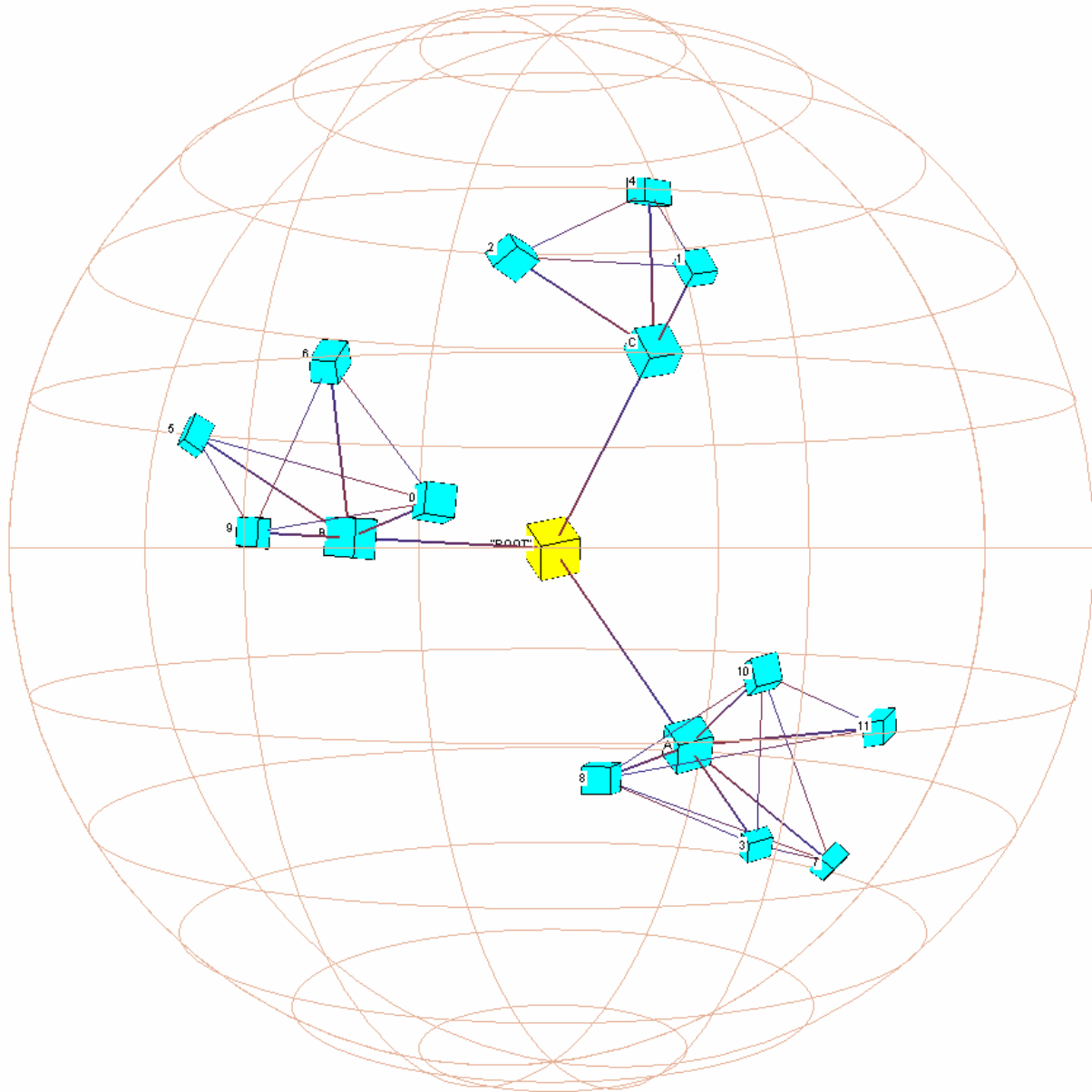


Figure 18: Clustered graph

5.4 Conclusions

Presented prototype has identified feasible mitigation plans for major visualization project risks. It validated usability of third party visualization framework for implementation of major user interface requirements. The prototype also identified major tailoring efforts necessary for development of required system capabilities.

Future prototyping effort will be targeting the following project risks:

- item relationship generation algorithm produces results that are inconsistent with client's expectations
- collection structure tree view produced by clustering of item relationship graph does not satisfy the client
- storage of intermediate results on a remote database will make the network bandwidth a bottleneck for system efficiency
- introducing local MySQL installation as a deployment requirement will limit usability of the system
- large effort necessary to design custom efficient data structures for storage of normalized usage records and produced item relationships.
- underestimated effort necessary to tailor the 3rd party visualization library for providing additional visual information
- inability of 3rd party visualization library to scale up to projected collection growth

According to the risks identified above the following items will be prototyping priorities:

- demonstrate item relationship generation from a sample usage log file of size and contents that could be manually examined and compared to produced results
- provide a larger example of producing graph structure tree using clustering
- demonstrate usage of server-less SQL based database for storage of intermediate results
- demonstrate modified 3rd party visualization tool with a side bar for displaying node information
- demonstrate semi-automated graph clustering and creation of a structure tree for a larger sample graph
- provide for client a visualization example where not all nodes of original graph are presented in the structure tree, but only nodes representing sub-graphs with granularity above certain level (e.g. 100 items per sub-graph node).

6. Glossary for Domain Description

Clustered image

A group of image sharing same attributes

Data mining

Collecting and retrieving meaningful data from a large database for solving problems

Data normalization

Techniques on eliminating data redundancy and dependency

Development stakeholders

Stakeholders of a software system that involve in actual development of the software

Digital archive

Collection of electronic and multi-media documents, such as images, photos, audios, and videos

Digital library

A library that houses electronic and multi-media documents which are accessible through a public terminal

Documents impact ranking

Measure and rank the impact that a document has relative to other documents

Graphical user interface (GUI)

An interface that contains graphics, such as images and photos

H3Viewer

An open-source package that can visualize a set of data in tree format.

IKIWISI

The “I know it when I see it” syndrome.

Independent verification and validation

A type of role in software development who test and verify the product before release

Information Data Management Department (IDM)

A division under USC Information Service Division (USC-ISD) that is responsible for managing information systems and databases.

Initial spiral

First phrase of the spiral model of software development, such as managing requirement objectives, constraints, alternatives, and risks

Log data

A type of record showing who access the data

Los Alamos National Laboratory (LANL) Research Library (RL)

A research library currently uses a recommender services that is using algorithms similar to the proposed new analysis system for USC Digital Archive.

MBASE

Model-based Architecting and Software Engineering, focuses on ensuring that a project's product models, process models, property models, and success models are consistent and mutually enforcing

MySQL

An open-source SQL (structured query language) database.

NASA Technical Reports Archive

A research archive currently uses a recommender services that is using algorithms similar to the proposed new analysis system for USC Digital Archive.

Open-source

Free packages that are written by other people

Open Video Project

An archive currently uses a recommender services that is using algorithms similar to the proposed new analysis system for USC Digital Archive.

Operational stakeholders

Stakeholders of a software system that involve in actual uses of the software

Prototype

Sketching of system interfaces

Recommender service

A system that provide library users a set of relevant documents based on their previous document searches

SQL

Structured Query Language; allow the program to retrieve a specific set of data from a database.

Usage trend

Typical pattern (behavior) for people who use the system

Usage analysis system

A system that analyze usage trend and generate results based on various analyses

USC Digital Library

USC online library catalogue that enable USC students to search for materials.

USC Digital Archive

Part of the USC Digital Library system which houses electronic materials (book, journals, picture, streaming media, etc.)

USC Information Services Division (USC-ISD)

USC organization that is responsible for computer networking, library services, academic computing, and telecommunications.

WinWin

Develop software and system requirements, and architectural solutions, as win conditions negotiated among a project's stakeholders.

7. Appendices

USC Information Services Division—Policies Governing the Use of Computing Resources at USC

<http://www.usc.edu/isd/policies/computing/>

Sample log data files

KEY	CLIENTIP	VIEWMOMENT	OBJECTNAME	SESSIONID
1	127.0.0.1	06-Jan-2005 04:13:59 PM	acsc-m305	
	BF129CE2D2167CF827D5C0B4582D6A31			
2	127.0.0.1	06-Jan-2005 04:14:16 PM	acsc-m1146	
	BF129CE2D2167CF827D5C0B4582D6A31			
3	127.0.0.1	06-Jan-2005 04:14:23 PM	acsc-m305	
	BF129CE2D2167CF827D5C0B4582D6A31			
4	128.125.65.204	06-Jan-2005 04:22:52 PM	bhe-m11	
	CD7B5D7EBF517AAE7A9C24B0E38B5A6A			
5	68.233.230.255	09-Jan-2005 08:40:24 PM	acsc-m1250	
	A28237096A8C56F4A5429FCDA5F1DF1D			
6	68.233.230.255	09-Jan-2005 08:42:05 PM	acsc-m989	
	A28237096A8C56F4A5429FCDA5F1DF1D			
7	68.233.230.255	09-Jan-2005 08:42:15 PM	acsc-m1250	
	A28237096A8C56F4A5429FCDA5F1DF1D			
8	68.233.230.255	09-Jan-2005 08:42:29 PM	acsc-m989	
	A28237096A8C56F4A5429FCDA5F1DF1D			
9	68.233.230.255	09-Jan-2005 08:42:32 PM	acsc-m989	
	A28237096A8C56F4A5429FCDA5F1DF1D			
10	68.233.230.255	09-Jan-2005 08:42:45 PM	acsc-m990	
	A28237096A8C56F4A5429FCDA5F1DF1D			
11	68.233.230.255	09-Jan-2005 08:42:55 PM	acsc-m982	
	A28237096A8C56F4A5429FCDA5F1DF1D			
12	68.233.230.255	09-Jan-2005 08:43:11 PM	acsc-m989	
	A28237096A8C56F4A5429FCDA5F1DF1D			
13	68.233.230.255	09-Jan-2005 08:43:43 PM	acsc-m989	
	A28237096A8C56F4A5429FCDA5F1DF1D			
14	68.233.230.255	09-Jan-2005 08:44:01 PM	acsc-m989	
	A28237096A8C56F4A5429FCDA5F1DF1D			
15	68.233.230.255	09-Jan-2005 08:44:08 PM	acsc-m989	
	A28237096A8C56F4A5429FCDA5F1DF1D			
16	68.233.230.255	09-Jan-2005 08:44:20 PM	acsc-m990	
	A28237096A8C56F4A5429FCDA5F1DF1D			
17	68.233.230.255	09-Jan-2005 08:45:19 PM	acsc-m990	
	A28237096A8C56F4A5429FCDA5F1DF1D			

18	68.233.230.255	09-Jan-2005 08:46:45 PM	acsc-m990
A28237096A8C56F4A5429FCDA5F1DF1D			
19	68.233.230.255	09-Jan-2005 08:48:08 PM	wpamaps-m10
A28237096A8C56F4A5429FCDA5F1DF1D			
20	68.233.230.255	09-Jan-2005 08:48:16 PM	wpamaps-m10
A28237096A8C56F4A5429FCDA5F1DF1D			
21	68.233.230.255	09-Jan-2005 08:49:27 PM	wpamaps-m11
A28237096A8C56F4A5429FCDA5F1DF1D			
22	68.233.230.255	09-Jan-2005 08:49:45 PM	wpamaps-m11
A28237096A8C56F4A5429FCDA5F1DF1D			
23	68.233.230.255	09-Jan-2005 08:49:55 PM	wpamaps-m12
A28237096A8C56F4A5429FCDA5F1DF1D			
24	68.233.230.255	09-Jan-2005 08:55:19 PM	acsc-m989
36A40B70BA8CF6F2DB00A6D13BD3720E			
25	68.233.230.255	09-Jan-2005 08:55:28 PM	acsc-m989
36A40B70BA8CF6F2DB00A6D13BD3720E			
26	68.233.230.255	09-Jan-2005 08:55:41 PM	acsc-m989
36A40B70BA8CF6F2DB00A6D13BD3720E			
27	68.233.230.255	09-Jan-2005 08:56:43 PM	wpamaps-m10
36A40B70BA8CF6F2DB00A6D13BD3720E			
41	128.125.65.204	14-Jan-2005 11:08:36 AM	chinese-m786
93D784AA72BCB4A8F632B346019A76A3			
42	63.199.206.182	14-Jan-2005 11:48:03 AM	bhe-m11
066C8CAEF43954946FDDDA4C7FBB77			
43	128.138.6.200	14-Jan-2005 11:48:21 AM	chinese-m36
673392701B79B23097805CABFD5271B1			
44	63.199.206.182	14-Jan-2005 11:48:26 AM	acsc-m20
066C8CAEF43954946FDDDA4C7FBB77			
45	63.199.206.182	14-Jan-2005 11:48:26 AM	acsc-m20
066C8CAEF43954946FDDDA4C7FBB77			
46	63.199.206.182	14-Jan-2005 11:48:33 AM	bhe-m28
066C8CAEF43954946FDDDA4C7FBB77			
47	63.199.206.182	14-Jan-2005 11:48:39 AM	bhe-m15
066C8CAEF43954946FDDDA4C7FBB77			
48	63.199.206.182	14-Jan-2005 11:48:44 AM	bhe-m13
066C8CAEF43954946FDDDA4C7FBB77			
49	128.138.6.200	14-Jan-2005 11:51:21 AM	chinese-m1038
673392701B79B23097805CABFD5271B1			
50	151.201.154.68	14-Jan-2005 12:06:17 AM	kada-m1941
4E1D54ECFD3C2FCC440A96CDC05BE049			
51	67.160.133.185	14-Jan-2005 01:36:02 PM	acsc-m305
EBFDC71B93419E55996D931D941A1DAD			
52	68.4.141.133	14-Jan-2005 02:03:04 PM	acsc-m1179
5A571C90DBD8550016786967709B8F18			
53	68.4.141.133	14-Jan-2005 02:03:40 PM	acsc-m1179
5A571C90DBD8550016786967709B8F18			
54	12.13.78.14	14-Jan-2005 02:09:10 PM	gg-m345
D97E40A911A748EBFE393959A0DF93B9			