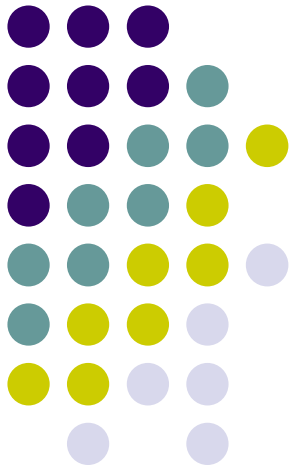
Data Mining of Digital Library Usage Data

- TRR ARB -

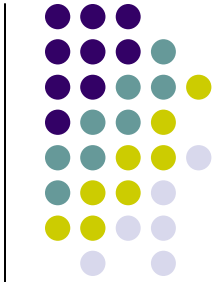
Client: Jewel Ward

Team 7

April 20th 2005



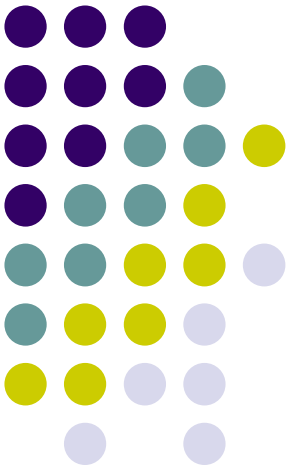
Presentation Outline



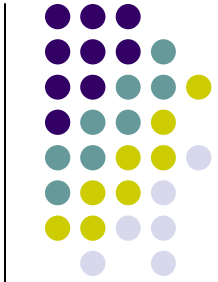
- Project Overview
- System Demonstration (business case)
- Requirements and Quality Assurance
- Transition plan
- Support Plan

Project Overview

Maks Krivokon

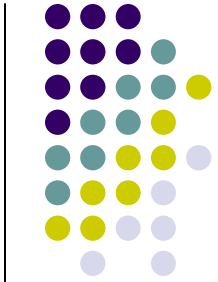


Overview



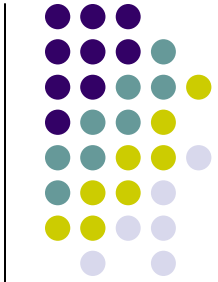
- How it all started
- What we have
- So what?

How it all started



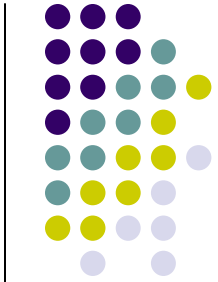
- Analysis
  - user based co-retrieval
  - session based analysis
- Visualization
  - 2D web-based interface
  - Visual thesaurus vs. H3Viewer
  - GUI proof of concept

What do we have?



- Analysis
  - Market Basket Analysis
  - hierarchical relationship clustering
  - clustering by collection
- Visualization
  - hyperbolic 3d view of the object tree
  - search by object name
  - file-system like browsing
  - view object info in browser window

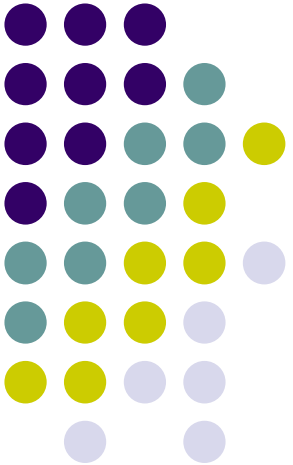
So what?



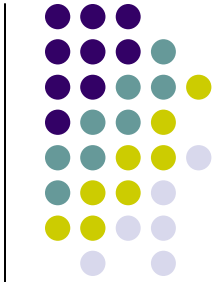
- What Digital Archive usage looks like?
- Which objects are important?
- Find good representatives of Greene & Green e collection?
- What can you tell about chs-m2682?
- What do “wpamaps” go with?

Requirements satisfaction & Quality Assurance

Bo Hye Lee

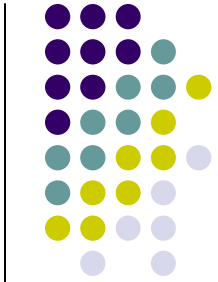


Outline



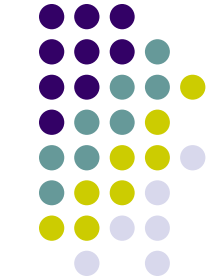
- (Old & New) Requirements Satisfied
- Requirements Dropped or Changed
- Quality Assurance
  - Peer Review
  - Requirement-based Testing
  - Requirement-based Code-level Testing
- Summary
- Q&A

(Old & New) Requirements Satisfied



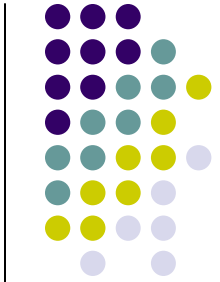
- Importing usage data from log file (SR-1)
- Relationship Generation (SR-3)
- Generate collection structure tree (SR-4)
- Visualization (SR-6)
- Omit mal-formatted retrieval records (SR-7)
- Omit usage log file with 0 valid records (SR-8)
- Do not visualize analysis report of corrupted format (SR-10)
- **NEW last-minute REQ.**: Search an object node in the visualized graph, focus and highlight the node
- **NEW last-minute REQ.**: Display the information on the selected or searched node.

Requirements Dropped or Changed



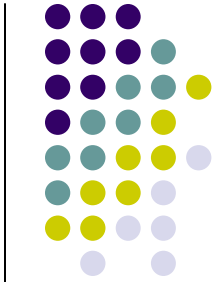
- Dropped: Remove usage data imported from the specified log file (SR-2)
  - > Removing reports will be done using windows file system.
- Changed: Graph node statistics (SR-5)
  - > Not the node statistics but the information on the node selected (Searched).
- Dropped: Do not create analysis report in case of abortion (SR-9)
  - > System not using DB due to storage constraint.
- Dropped: Remove analysis report (SR-11)
  - > Removing reports will be done using windows file system.

Quality Assurance



- Peer Review
- Requirement-based testing
  - Performance testing
- Intensive requirement-based code level testing
  - Performance testing
  - Concurrency
  - Scalability
  - Usability

QA: Peer Review

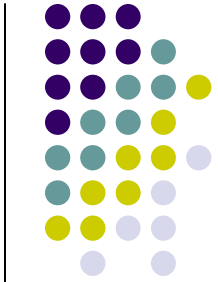


- By other team members and IV&Vers
- On all the documents
- Summary of Peer Review data*

Time meeting effort (total time * people)	18.5 * 3 = 55.5
Total defects found	111
Defects asserted per minute of meeting	111/55.5 = 2
Peer Review Effort	27.5 + 55.5 + 20 = 103
Defect Removal Effectiveness (total defects found / effort)	111/103 = 1.07
Defect Density	111/394 = 0.28
Peer Review Rate	394/55.5 = 7.10

*Peer Review statistics from Peer Review Report v.1.0

QA: Req.-based Testing



- By IV&Vers
- On the build as of April 3rd, 2005
- To ensure the core capabilities are implemented and working correctly
- Major three requirement-based functionality groups. (Test Plan & Cases)
  - Relationship generation with input/output
  - Clustering
  - Visualization

QA: Req.-based Testing (Cont'd)

- Performance testing and the results*

	Input data	Relationship Generation	Visualization
TI:TC**	TI-01:TC-01-01	TI-03: TC-03-03	TI-05:TC-05-03, TI-06:TC-06-01, TC-06-02
Result	Pass	Pass	Pass

- Some of TIs and TCs were dropped according to the requirement changes

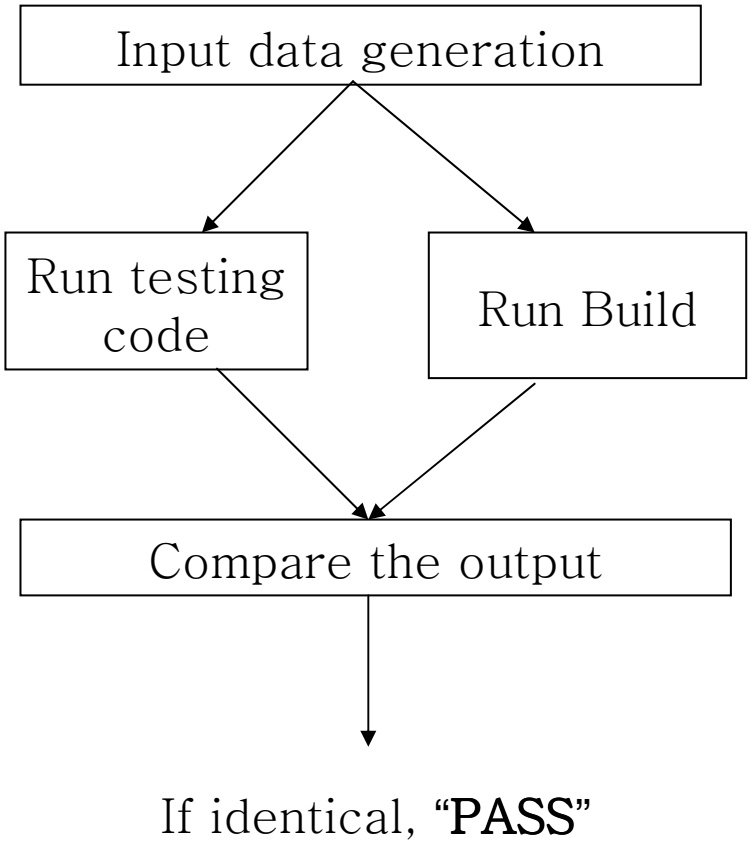
* Refer to Test Procedure and Results and Regression test
** TI: Test Identifier / TC: Test Case

QA: Req.-Based Code Level testing

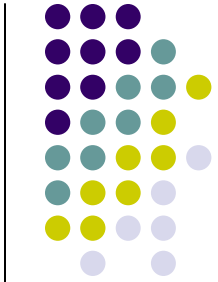
- By team members
- On unofficial builds.
- To ensure the level of services for relationship generation.
- Testing for:
  - Scalability(LS-5): Input data size
  - System dependability(LS-1): Input data format control
  - Performance: Output validation

QA: Req.-based Code-level testing (cont'd)

- Process
  - **Input data generation**: Manually / by automation
  - **Testing Code**: Another complete implementation of Relationship generation
  - **Build**: unofficial internal source code
  - **Comparison**: Pick a certain number of nodes in one output and look them up in the other, if identical, then "PASS"



QA: Req.-based Code-level testing (cont'd)



- Scalability: Data Size test cases and the results

Input size (entry)	0	1	100	500	1000	5000	10000	50000	100000	150000
Testing Code	S	S	S	S	S	S	S	S	S	S
Source Code	S	S	S	S	S	S	S	S	S	S
Result	P	P	P	P	P	P	P	P	P	P

S: Successfully Run / N: Not successfully Run / P: Pass / F: Fail

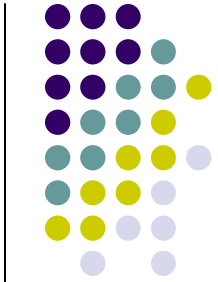
QA: Req.-based Code-level testing (cont'd)

- Reliability: Irregular data control test cases and the results

Test cases	All same ids, in one big session	No relationship in 100* entry input	No relationship in 500* entry input
Testing Code	No relationship matrix generated	No relationship matrix generated	No relationship matrix generated
Source Code	No relationship matrix generated	No relationship matrix generated	No relationship matrix generated
Result	Pass	Pass	Pass

* 100 and 500 are the numbers found on formal req.-based testing by IV&Vers

Summary



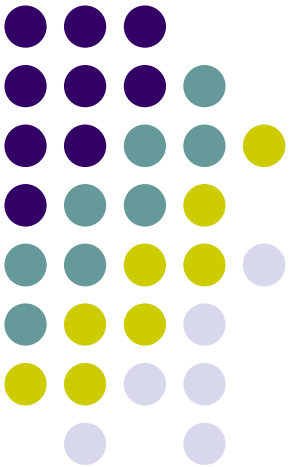
- Most of the requirements plus new last-minute requirements are satisfied
- The requirements dropped have alternative ways to fulfilled.
- The requirements changed has a reason for better goal.
- Good amount of effort for Quality Assurance has been made.
- The core capabilities were tested and passed.
- Extensive testing has been made to ensure the level of services.

Requirements satisfaction & Quality Assurance

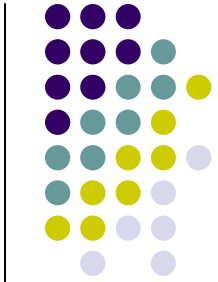
Questions?

Summary of Transition Plan

Genesan Kim

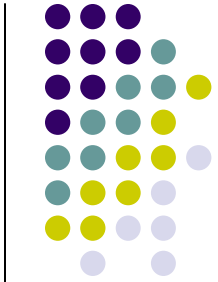


Outline



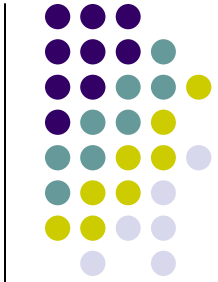
- HW & SW
- Staff Preparation
- Operational testing, training & evaluation
- Required Resources
- Transition Milestones
- Source code & Documentation

HW & SW



- PC Computer, 50MB Hard disk, 128 MB RAM
- Linux Operating System
- No site preparation – Client expecting a Linux server on site
- No hardware installation required
- SW installation – copy built system over to target system

Staff Preparation

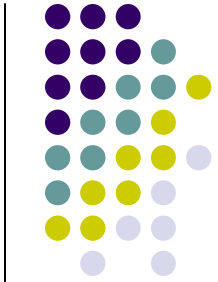


- Training deliverable – User Manual
- 2hrs per training session – involves walkthrough of system with user manual, time for independent use of system + feedback
- Training sessions for client and maintainer
- Session will involve training of all capabilities of the system
- Measure of success depends on client ability to go through each capability with the new system after training sessions

Operational testing, training & evaluation

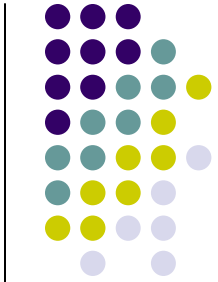
- Training in all of system's capabilities and requirements
- Step by step procedure, will assure quality and understanding after each step
- Session will involve client feedback and evaluation of the system
  - System should improve efficiency in analyzing usage log data compared to manual analysis
  - User friendly interface
  - All requirements satisfied, client expectations met
- 2 training sessions for client, 1 for maintainer
  - Potential users trained by client

Required Resources



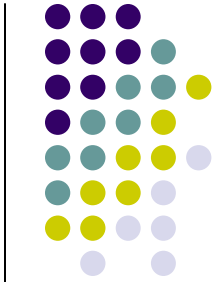
- Trainer preparation time = 4 hours. Since training preparation will be done by the development team, cost = $0.
  - Client will be trained as an administrator of the system, and will be given instructions on how to analyze log data using the system and distribute results of the analysis to regular users. Client training time = 2 hours; Average client salary = $35 per hour. Therefore cost = 2 * $35 = $70.
  - Maintainer training time = 2 hours; Average maintainer salary = $30 per hour. Cost = 2 * $30 = $60.
  - Total training costs = $70 + $60 = $130
- Data Preparation = $0, data provided by DA system
- COTS licenses = $0

Cont..



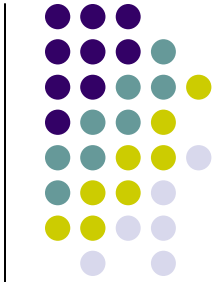
- Operational readiness testing:
  - This testing will be done by the client through comprehensive evaluation of major functionality of the system. Estimated testing time = 2 hours; Average client's salary = $35. Cost = 2 * 35 = $70.
    Total testing costs = $70
- Site preparation:
  - The proposed system is a standalone application which can be installed on any workstation running Linux, which does not require any special preparations
    Site preparation costs = $0
- Facilities preparation:
  - Proposed system does not require use of any specialized facilities.
    Facility preparation costs = $0
- Equipment purchase:
  - The developed system will be installed on existing Digital Library equipment and also on personal workstation of the client; therefore no equipment will be purchased.
    Equipment cost: = $0

Transition Milestones



- Customer product review – 4/21/05
- Documents completion – 4/29/05
- Training – 5/2/05
- Software and documents delivered to the customer – 5/4/05
- Project closeout – 5/4/05

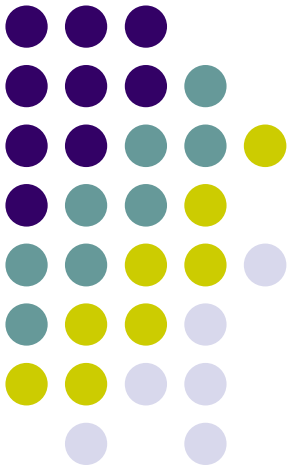
Source code & Documentation



- Source code will include extensive comments
- Documentation will be supplied

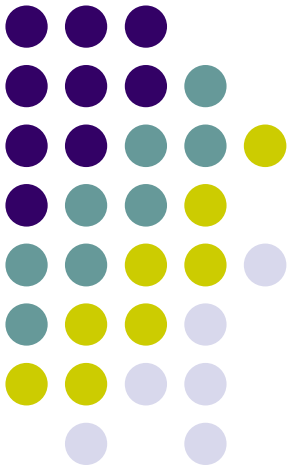
Transition Plan

Questions?

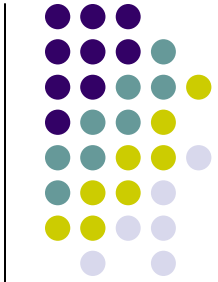


Support Plan

Vu Nguyen

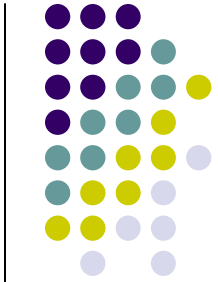


Support Plan



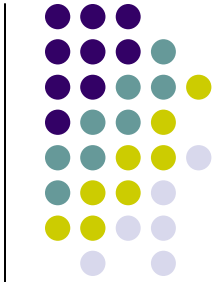
- Responsibilities
  - Support stage is handled by USC ISD.
    - Representative: Jewel Ward – responsible for maintenance strategies
    - Maintainer: Jeff Pearson – responsible for maintenance and development
- Estimated effort: 240 hours/year

Support Plan



- Strategies
  - Estimated lifetime: 1 – 2 years
  - Minor releases
    - Fixing bugs, including minor changes
    - Synchronizing with H3Viewer and MCL components
    - Each release may consist of activities requirements, A&D, test and release
    - Schedule: up to 2 months
  - Major releases
    - Incorporating major changes, new requirements:
      - New solutions for analyzing log (different implementation)
      - Accommodate collection growth
      - Support other different kind of data
    - Each release may consist of phases: inception, elaboration, construction and transition
    - Schedule: from 2 months

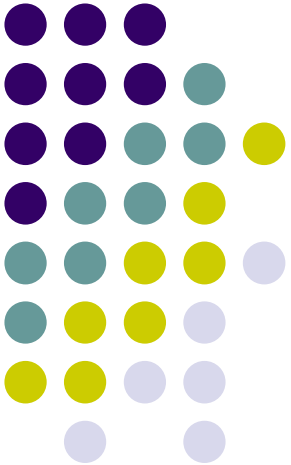
Support Plan

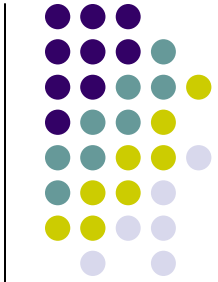


- Environment
  - Hardware: a computer with minimum configuration
    - RAM: 128 MB. Recommended: 256 MB
    - HDD: 50 MB free space. Recommended: 100 MB
  - Software:
    - OS: Linux 2.6.9
    - Code editor and compiler: Anjuta
- Support documents
  - OCD, SSRD, SSAD, FRD, etc
  - Support Plan, Test Plan, Regression test package, User Manual, Training Materials.

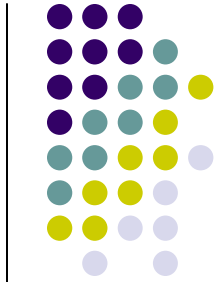
Support Plan

Questions?





Any Comments?



Thank you!