

**PATTERNS IN THE DAILY DIARY OF THE 41st PRESIDENT,
GEORGE BUSH**

A Thesis

by

SHREYAS KUMAR

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2005

Major Subject: Computer Science

**PATTERNS IN THE DAILY DIARY OF THE 41st PRESIDENT,
GEORGE BUSH**

A Thesis

by

SHREYAS KUMAR

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,
Committee Members,
Head of Department,

Frank M. Shipman, III
Richard K. Furuta
Lauren Cifuentes
Valerie E. Taylor

December 2005

Major Subject: Computer Science

ABSTRACT

Patterns in the Daily Diary of the 41st President,
George Bush. (December 2005)

Shreyas Kumar, B. Arch., I.I.T. Roorkee

Chair of Advisory Committee: Dr. Frank M. Shipman, III

This thesis explores interfaces for locating and comprehending patterns among time-based materials in digital libraries. Time-based digital library materials are like other digital library materials in that they are comprised of data and metadata. In addition, they have a time or period of time attached to each data item. The specific focus of this thesis is on fine-granularity items – items that have relatively little data and cover brief periods of time. In such a context, people often are left to discern patterns of activity by retrospectively making sense of the collection or parts thereof. The specific domain chosen for the implementation is the daily diary of President George Bush, the 41st president of the USA. This project developed a searching and browsing interface, which allows people to study the relationship between activities and people in the library data. As part of this thesis, a corpus of the Presidential daily diary was digitized. Two interfaces were provided to this corpus, one based on a standard information retrieval engine (Greenstone) and another presenting time-based visualizations of data items. An evaluation was conducted to explore the relative strengths and weaknesses of these two interfaces.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere gratitude to those who helped, guided or supported me in this challenging yet interesting endeavor.

I am very grateful to Dr. Frank Shipman, the chairman of my thesis advisory committee, for his guidance and support all the way. I am also very thankful to my committee members, Dr. Richard Furuta and Dr. Lauren Cifuentes for their valuable suggestions and feedback.

I would also like to acknowledge:

Dr. Nancy Amato, Professor and Graduate Advisor and Dr. Bart Childs, Professor and former Graduate Advisor, Department of Computer Science for their advice throughout my graduate studies,

Dr. Valerie Taylor, Chair of the Department of Computer Science, Texas A&M University for being very supportive,

The Rotary Foundation of the Rotary International, IL, USA for partially supporting my graduate education through the Multi-year Rotary Ambassadorial Scholarship,

Mr. William Douglas Moore for his encouragement and mentorship throughout my stay in College Station,

Dr. Charles Hermann, Associate Dean for Academic Affairs, The Bush School of Government and Public Services, Texas A&M University for his encouragement,

Dr. Jeffrey A. Engel, Assistant Professor, Bush School of of Government and Public Services, Texas A&M University for his useful suggestions and feedback,

Dr. Edward Douglas Menarchik, Assistant Administrator for Policy & Program Coordination, US Agency for International Development, Washington D.C., and former Director of George Bush Presidential Library and Museum, College Station for facilitating access to the Presidential library,

Dr. Robert Holzweiss and Mr. Gregory Mouchen, George Bush Presidential Library and Museum, College Station for providing copies of a section of the daily diary of President Bush and for valuable suggestions throughout the project,

Ms. Linda Edwards, Director of International Outreach, Texas A&M University for being a very inspirational and caring boss at work,

My fellow researchers at the Center of Digital Libraries, Dr. Luis Francisco-Revilla, Dr. Haowei Hsieh, Michael Moore, Raghu Akkapeddi, Rajiv Badi, Unmil Karadkar, Anna Zacchi, Sarah Davis and Yi Yang for their suggestions and help,

Ms. Viki Colson, and other staff of the IRB at Texas A&M University for guiding me through the processes of Research compliance,

The Thesis Office at Texas A&M University for their careful review of my thesis document,

Members of the Greenstone Project, University of Waikato, New Zealand for developing a great open source tool,

All the anonymous respondents of the evaluation survey for their time and feedback,

My parents, Dr. Mridula Kumari and Mr. Shailendra Nath Sahu for their guidance, encouragement and support,

Harshita for being the best sister in the world, and

My wife Swapnil for her unconditional love.

TABLE OF CONTENTS

	Page
ABSTRACT.....	iii
ACKNOWLEDGEMENTS.....	iv
TABLE OF CONTENTS.....	vii
LIST OF FIGURES.....	ix
LIST OF TABLES.....	x
I INTRODUCTION.....	1
II PROBLEM STATEMENT AND DOMAIN.....	3
III RESEARCH GOALS.....	6
Access to Information.....	6
Pattern Finding.....	7
Visualization.....	7
IV RELATED WORK.....	9
V DEVELOPMENT OF THE DIGITAL LIBRARY.....	12
The Corpus.....	12
Planning.....	12
Scanning.....	13
Optical Character Recognition (OCR).....	14
Greenstone Repository Creation.....	16
VI VISUALIZATION.....	24
System Design.....	24
Tools and Technology.....	25
DigiLine.....	26
VII EVALUATION.....	31

	Page
Results.....	32
VIII DISCUSSION.....	35
IX CONCLUSION.....	36
REFERENCES.....	38
APPENDIX A.....	41
APPENDIX B.....	42
APPENDIX C.....	43
VITA.....	44

LIST OF FIGURES

	Page
Figure 1 Scanning.....	13
Figure 2 Optical Character Recognition.....	15
Figure 3 Greenstone Librarian Interface – Gather	18
Figure 4 Greenstone Librarian Interface - Enrich.....	19
Figure 5 Greenstone Librarian Interface - Design.....	20
Figure 6 Greenstone Interface - Search Page	22
Figure 7 Greenstone Interface - Result Page	23
Figure 8 System Overview.....	25
Figure 9 DigiLine Search Interface on Left with Scanned Document on Right. ..	27
Figure 10 DigiLine Results are Shown on Right with Timeline Below.....	28
Figure 11 DigiLine with a Timeline Displaying Two Different Search Results	29
Figure 12 DigiLine after Navigation to Document from Timeline.....	30

LIST OF TABLES

	Page
Table 1 OCR Quality Testing.....	16
Table 2 Survey Results.....	32

I INTRODUCTION

Digital libraries have large amounts of data and it can be difficult for a user to find relevant material from the corpus of materials. Information retrieval and general search technologies make it possible to locate resources within a library that have certain characteristics, but there are occasions when the information of interest are the relationships between the elements, rather than the library elements themselves. One class of content where the relationships between elements are more important than the content of the individual elements is fine-grained time-based elements.

Time-based digital library materials are like other digital library materials in that they are comprised of data and metadata. The one requirement is that they have a time or period of time attached to each data item. The specific focus of this thesis is on fine-granularity items – items that have relatively little data and cover brief periods of time. In such a context, people often are left to discern patterns of activity by retrospectively making sense of the collection or parts thereof. Time-based digital library materials offer interesting possibilities to manipulate the browsing, because the data is arranged on a timeline. This thesis explores interfaces for locating and comprehending patterns among time-based materials in digital libraries.

Time-based digital library materials offer interesting possibilities to manipulate the browsing, because the data is arranged on a timeline. This thesis explores interfaces for locating and comprehending patterns among time-based materials in Digital Libraries.

This project is focused on the design and development of an interface that enables access to time-based digital library materials that combines searching, browsing and visualization of information. More specifically, the goal is to develop a web-based interface, which can search the digital library to find the occurrences of people, occurrences of events and the correlation between the various entities with respect to time. The chosen domain for study is the daily diary of President George Bush, as it contains interesting data and the data is stored date-wise. Also, the diary is an interesting tool for research for researchers for many domains, including History and Political Science.

The next section describes the problem and domain in greater detail. Following this is a description of research goals and related work.

II PROBLEM STATEMENT AND DOMAIN

With the development in computing power and drop in the prices of storage media, there is a rapid rise in the amount of data people generate in daily life. Small artifacts that were previously ephemeral, such as daily schedules, are now being archived. With the development in scanning and character recognition tools, large amounts of older documents and records concerning schedules are also being digitized. The enormous amount of data generated poses a challenge for its effective use in a meaningful fashion.

Researchers often want to identify patterns of activity and relationships between scheduled activities. This is particularly relevant in terms of events and people of broad interest as identification of patterns and relationships between data can generate valuable analytical results for historians, political scientists, etc. These results can be of historical or strategic value.

After a President of the United States of America leaves office, the papers and memos generated during his administration are archived in presidential libraries. The George Bush Presidential Library and Museum is located on the campus of Texas A&M University in College Station, Texas. As time passes on, some of the documents are declassified and made available to the general public. Recently, in 2004, parts of the daily diary of the 41st U.S. President George Bush were

declassified. These records contain a detailed account of the activities and meetings of the former President. Currently, the documents are in paper format and are stored in the George Bush Presidential Library in College Station, Texas. The diary runs into thousands of pages and it is a cumbersome task to physically go to the library and locate specific information. Yet this is exactly what Presidential historians and political scientists are currently left to do, when conducting research.

The Presidential Diary is a detailed tabular account of the President's activity. It is used to document the President's actual activity during his term in office, rather than the planned activity. Examples of entries are:

“6:51 – 7:17, The President motored from the South Grounds to Andrews AFB, Maryland”, and

“1:54 – 1:57, The President met with:

Jack F. Kemp, Secretary of Housing and Urban Development (HUD)

Warren E. Buffet, Chairman and Chief Executive Officer, Hathaway Berkshire Corporation, Omaha, Nebraska”

The Presidential diary contains a large amount of information, as the President's daily activity is captured almost every minute of everyday he is in office.

Practically, this information is of interest to political scientists and historians – and the system results of this thesis may help improve their ability to perform this type of analysis. Thus, the Presidential diary is an excellent choice for prototyping tools for accessing and analyzing fine-grained time-based materials.

[Political Science Resources 2005] provides insights into usefulness of political sciences resources online.

The declassified diary is a public document, and is intended to be publicly available. This thesis does not adversely impact any privacy or security interests of any person, government or entity. All documents made available through this project are provided by the George Bush Presidential Library and Museum at College Station.

III RESEARCH GOALS

This thesis explores interfaces for locating and comprehending patterns among time-based materials in Digital Libraries. As such, the project included the design, development, and evaluation of an interface that enables access to time-based digital library materials that combines searching, browsing and visualization of information.

ACCESS TO INFORMATION

Currently, the daily diary of the 41st President George Bush is only available in the hard copy format at the George Bush Presidential Library. It was determined that a Web-based interface would be developed to allow remote access to the diary. This interface could potentially be made accessible to interested library patrons world-wide.

Before materials can be accessed online, they must be digitized. A secondary goal of this thesis is to digitize a portion of the Presidential diary of George Bush. Various methods for scanning and converting to text were explored to determine what cost-effective options are available that best serve these tasks.

PATTERN FINDING

The project aims to facilitate researchers, students, and others wishing to find patterns in the diary. This interface was specifically designed to support searching the digital library to find the occurrences of people, occurrences of events and the correlation between the various entities with respect to time. For instance, the system can help find out who were the key advisors to the President in the days leading to a particular event.

For example, consider that a student of foreign policy wants to explore if there is a pattern in the activities of President Bush before the fall of the Berlin wall. The proposed tool can help her find whether or not the President met with a particular group of people or talked to a particular person, say, Mr. Gorbachev, over the telephone more frequently than normal during the period in question. Another example use is to explore the patterns of meetings of the President in the weeks preceding the 1991 Gulf war.

VISUALIZATION

The results of the search are shown on a graphical timeline. The design should also enable a comparison between two entities along a timeline. The goal of visualization is to enable the users to identify possible correlations between the occurrences of people, places and events in the diary. To aid in this effort, the system generates time-based histograms summarizing the search results.

The effectiveness of the interface was measured by conducting a user evaluation study, where potential users would compare the interface with a standard digital library interface.

IV RELATED WORK

This project builds on prior work on computational diaries, time-based visualizations, and tools supporting historians and political scientists.

Previous work has been done on diaries of a personal nature. Diaries have been studied as family communication tools [Fleuriot et al. 1998] and as Knowledge Management tools [Kovalainen et al. 1998]. Hornbaek presents a comparative study in a diary experiment [Hornbaek 2004]. Blogging, the authoring of diary-like expositions and their publication on the Internet, is closely related to online diaries and a study by Nardi presents an ethnographic report on blogging as a form of journalism [Nardi et al. 2004]. Some of the other related work includes work on Global Digital Museum [Takahashi et al. 1998], Pattern augmented digital libraries [Goh and Leggett 2000], Compus [Fekete and Dufournaud 2000], Picnic [Hunter and Choudhury 2005]. Related work on analyzing and manipulating workspaces was also studied [Hsieh and Shipman 2002] and [Shipman et al. 2001].

In the area of time-based visualizations, various techniques have been developed for providing an overview of the whole timeline, i.e. a zoomed out view, zoomed-in view with parts of the hierarchy and display entities at various depths in a hierarchy [Kumar et al. 1998]. Karam [Karam 1994] studied the

visualization of software events using state machine diagrams to model events and activities. A prototype timeline generator was implemented.

Many of the timelines developed are text based [Brownstone 1994] and [Waldman 1994] although some timelines make use of a graphical presentation [Mann 1993]. While timelines are used to illustrate a particular phenomenon, say American History [Smithsonian Institute 1993], the studies of these visualizations have not explored their impact on people's ability to identify correlations between time-based entities that were previously unrecorded.

One project that does look at supporting the identification of correlations is Chieu's work on event extraction along a timeline [Chieu and Lee 2004]. This work attempts to establish a relation between events and people (G8 leaders). The work presented in this thesis explores the integration of the event-people correlation of Chieu with a visual timeline presentation similar to that by [Kumar et al. 1998].

Among other projects related to computational support for political science, the THOMAS database [The Thomas Online Database 2005] provides text-based interface for the archives of the US Library of Congress. The National Archives of UK [The National Archives Search the Archives Page 2005] provides users the

options to search for keywords and then displays the relevant scanned document in a web-based interface. Each image in the archive is indexed with keywords that are then used for searching and retrieving the images. However there is no option to search through the textual content of the archive, and there is no option to search on the basis of a timeline.

V DEVELOPMENT OF THE DIGITAL LIBRARY

THE CORPUS

The Presidential diary of 41st President Bush includes approximately 10 pages for each day. This leads to an estimated 14,600 pages for four years. For the purpose of this research, 500 pages of data were obtained from the George Bush Presidential Library Museum at College Station. These 500 pages correspond to the time period of November 1990 to January 1991. Initially, the data was in the form of typewritten pages of letter size.

PLANNING

The major challenge was to determine the most suitable scanning and character recognition technology to accomplish the task. Three major steps were identified to develop the digital library: Scanning, OCR and collection and index creation using Greenstone. There was a logistical limitation on how many pages could be digitized in a reasonable time frame. It was determined that 3 months of data, particularly the three months prior to the start of the Gulf War, would cover enough time to include interesting patterns of occurrences and changes in the Presidential activities. .

SCANNING

The pages were scanned and saved in PDF format. Each page was named as yyyyymmdd_n.pdf. The pages were stored in folders according to year and month. While scanning, pages containing appendices were ignored.

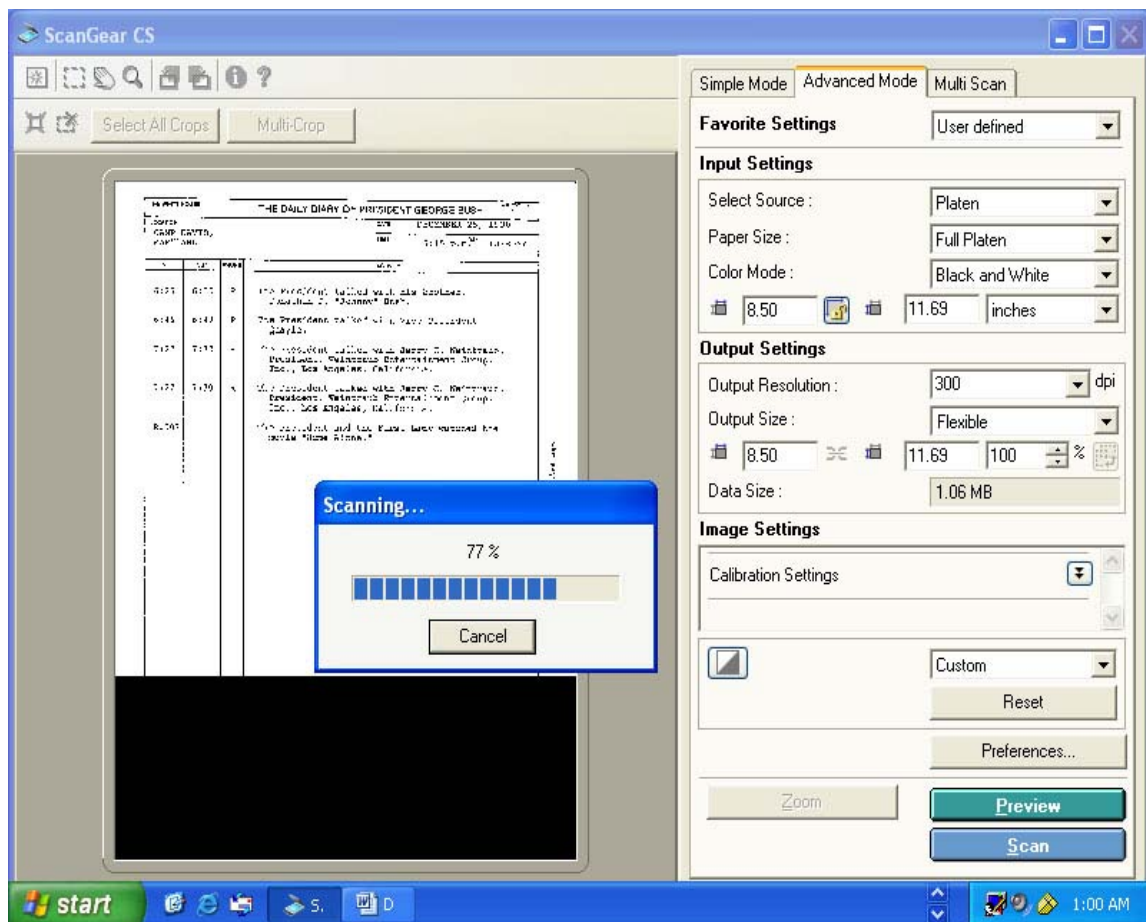


Figure 1: Scanning

Figure 1 shows the Scan Gear interface, with the settings, which were used in scanning the documents. Setting the color mode to Black and White was absolutely critical to this step. A default setting of color would lead to an enormous filesize – leading to a failure at the OCR stage.

OPTICAL CHARACTER RECOGNITION (OCR)

The Bush diary pages are typewritten pages in a tabular format. It was found that most OCR software has limitations when it comes to recognizing text in a tabular format. After various scanning and OCR software were evaluated, it was observed that Omni Page Pro software, provided reasonable accuracy. Also, it was found that, compared to simultaneous scanning and OCR, higher accuracy was achieved if the documents were first scanned and stored in PDF format, and then these PDF files were subjected to OCR. Even with the PDF files, the OCR process worked best with certain settings for generating the PDF files.

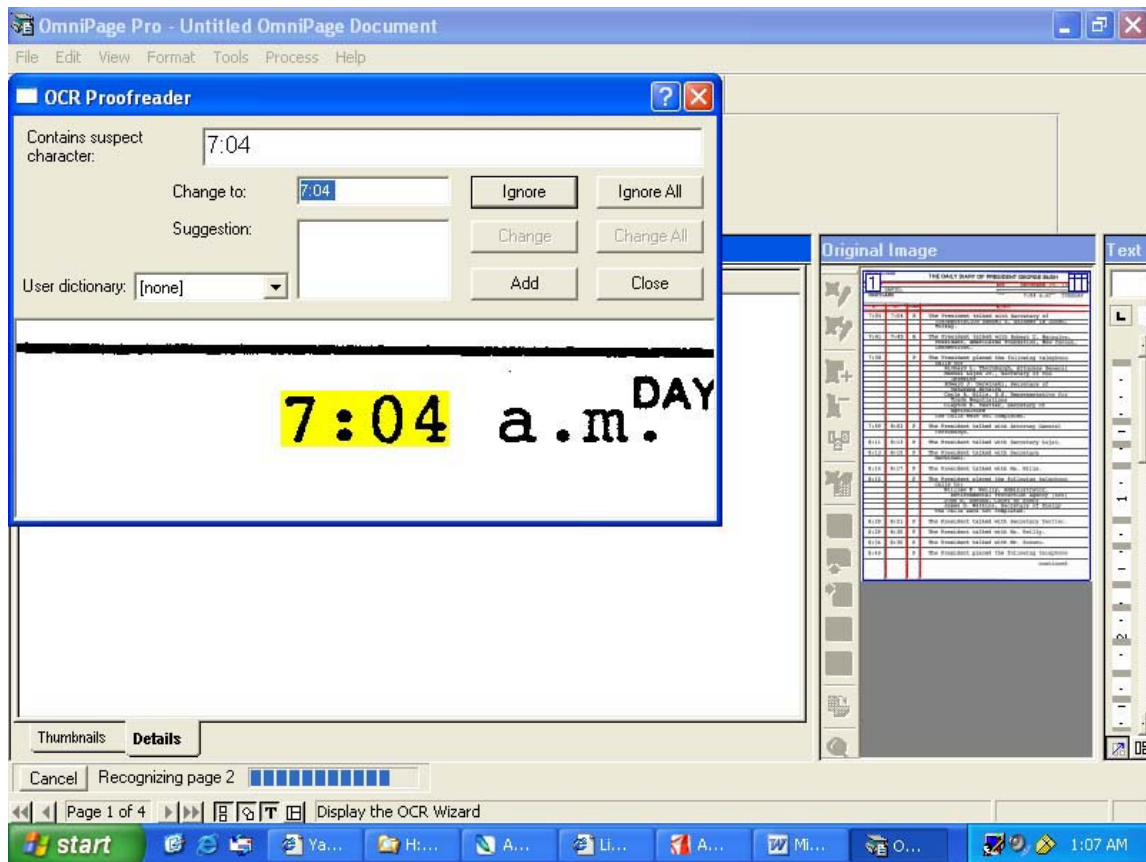


Figure 2: Optical Character Recognition

Figure 2 shows how the OCR process allows manual intervention in case of suspect characters. The suspect character is a character which could not be resolved by the OCR software. The character is highlighted by yellow and can be manually changed by the user.

Table 1 shows the quality of scanning and OCR when 2 software were tested for different scenarios.

Table 1: OCR Quality Testing

Products:	Read Iris Pro		Omni Page Pro	
	Plain text	Tables	Plain Text	Tables
OCR along with Scanning – in a single step	Good	Poor	Good	Poor
OCR from scanned PDF files (scanned as Color pages)	Fair	Poor	Fair	Poor
OCR from scanned PDF files (scanned as Black and White Pages)	Good	Poor	Good	Good

GREENSTONE REPOSITORY CREATION

Greenstone version 2.53 was used to support search of and access to library materials. Greenstone is a free open source set of digital library development tools and runtime software. It has a Librarian interface that provides the

functionality to create a digital library through a 5 step process: Download, Gather, Enrich, Design and Create. Download was not relevant for this project as the content was captured via scanning and OCR.

Gather

The first step is to gather the digital library contents. This process began by converting all the scanned images into html files, and named according to the date. For example, all pages (PDF files) related to November 1 1990, were converted to an html file named 1990_11_25.htm. The files were arranged in year-wise folders and subfolders for months. In this step, the root folder containing the entire year of content was added to the Greenstone repository. Figure 3 shows this process – the left pane allows browsing through the directory tree that can be included in Greenstone and the right pane shows the folders added to the collection. The left pane also allows for files to be downloaded from the internet.

Figure 3 shows the Gather step, where files can be added to the repository through Greenstone.

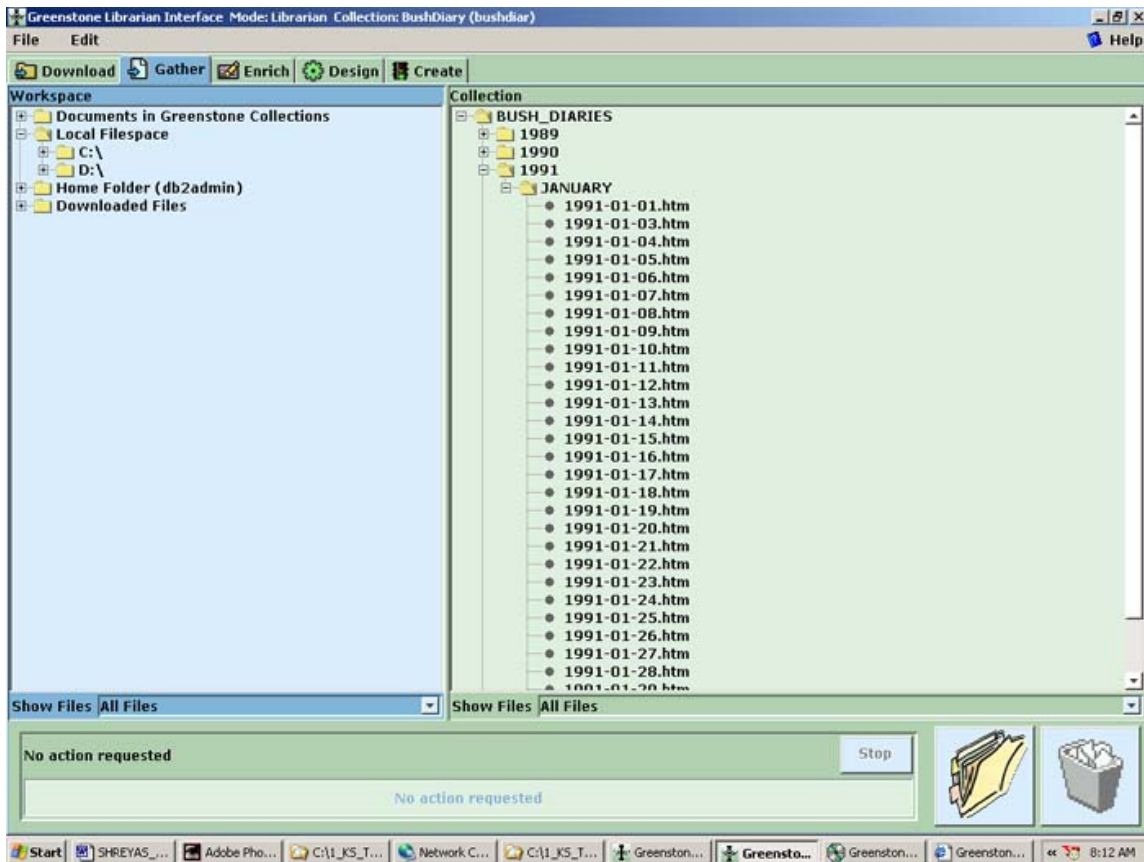


Figure 3: Greenstone Librarian Interface – Gather

Enrich (Indexing)

In the enrich phase, each page is indexed based on one or more parameters. The Bush Diary Collection is indexed on the basis of title, subject, and keywords. Since the date information is encoded in the title, the indexes facilitate searching on the basis of month or date. It was possible to index the pages with keywords that include the names appearing on that date, but that could lead to very high number of words getting indexed. Greenstone provides the option of browsing

through titles alphabetically, A-Z. Due to the representation of date in these titles, the alphabetic browsing mimics browsing by date. Figure 4 shows the indexing options.

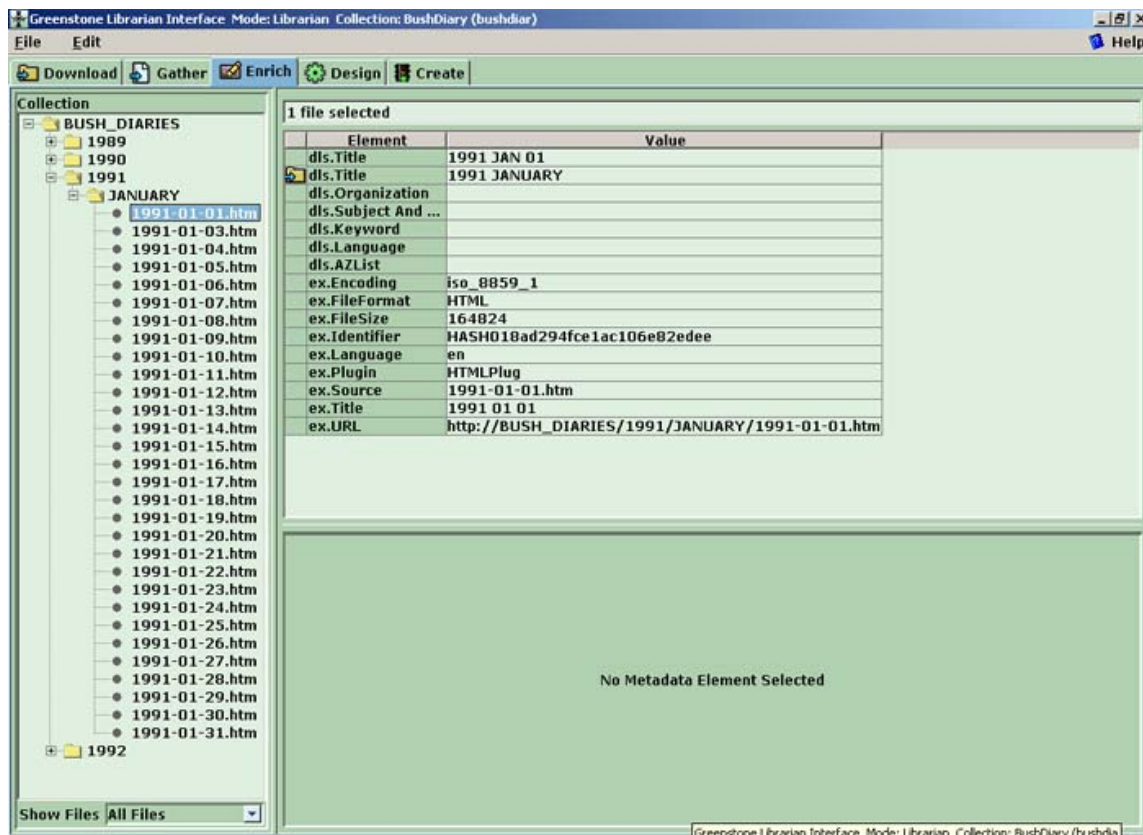


Figure 4: Greenstone Librarian Interface - Enrich

Design

The design section of the Librarian Interface allows the users to control different aspects of the appearance of the collection. An image uniquely identifying the

collection can be specified at this stage. Format features include commands that are used to change the appearance of the pages that are returned from the digital library. Additional options were document plug-ins, search types, cross-collection searches, partition indexes, browsing classifiers, translation text and metadata sets. Basic formatting commands and images were used to “brand” the Greenstone interface to the Presidential Diary collection.

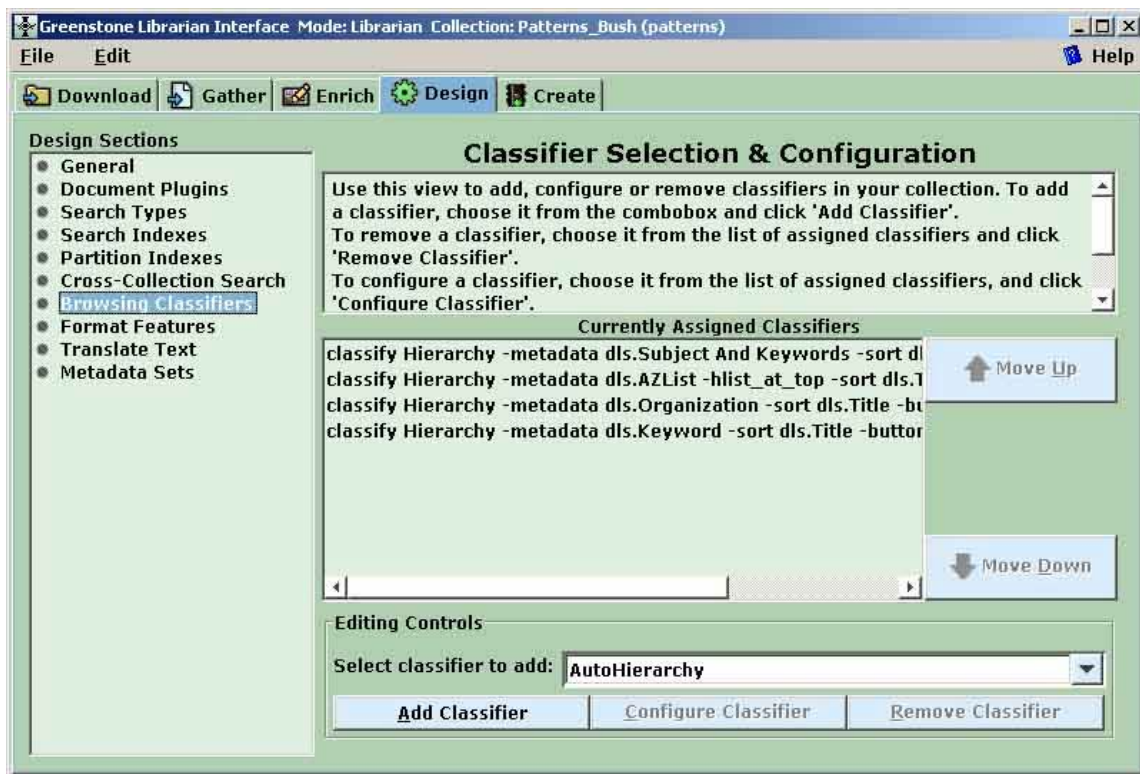


Figure 5: Greenstone Librarian Interface - Design

Figure 5 shows the classifier section & configuration. This is the last step before a collection is created.

Create

The create step includes assigns a unique alphanumeric hash code to each file included in the collection (e.g. HASH2c61b7f1bf6f5630cdd8ce). After a collection is created, it is ready for preview. After this step, the greenstone Librarian interface can be closed, and the collection can be viewed through the viewing interface, from the installed Greenstone homepage. Figure 6 shows the standard Greenstone-based interface to the Presidential Diary collection, which offers basic term-based search and title based browsing.

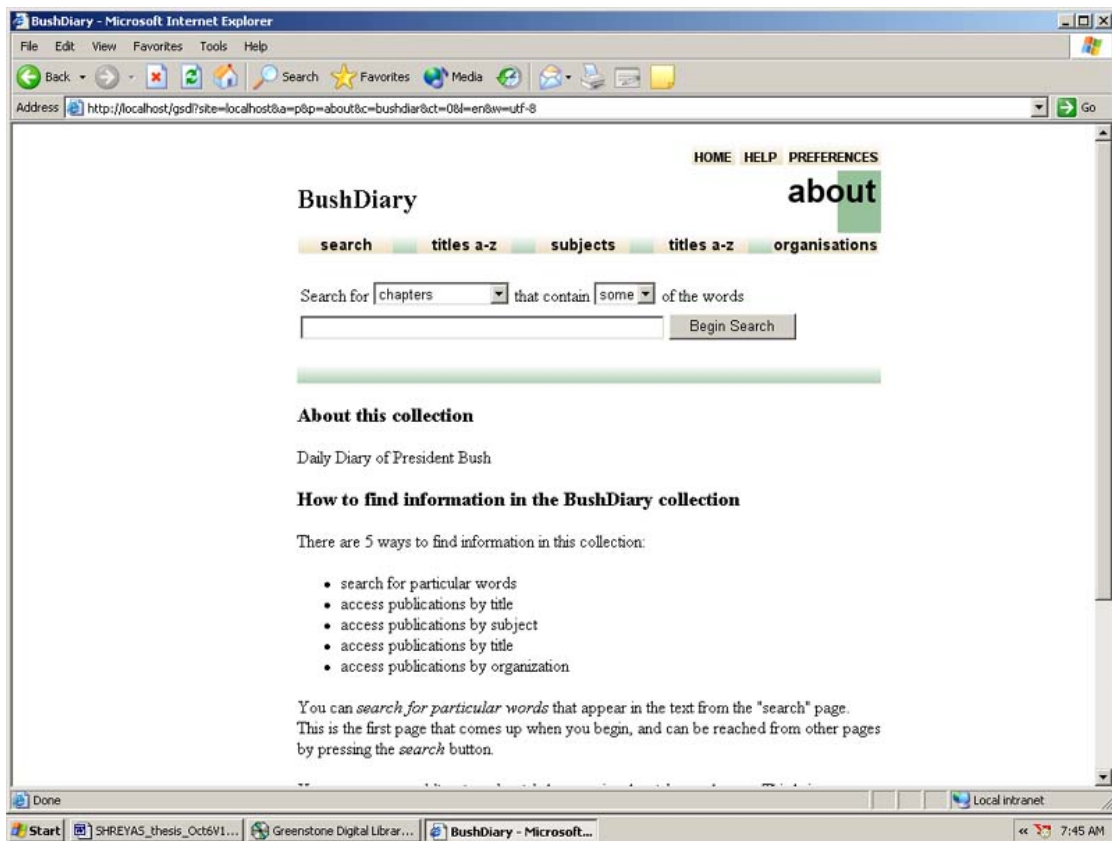


Figure 6: Greenstone Interface - Search Page

The standard Greenstone-based interface provides a list of page entries matching search terms as displayed in Figure 7.

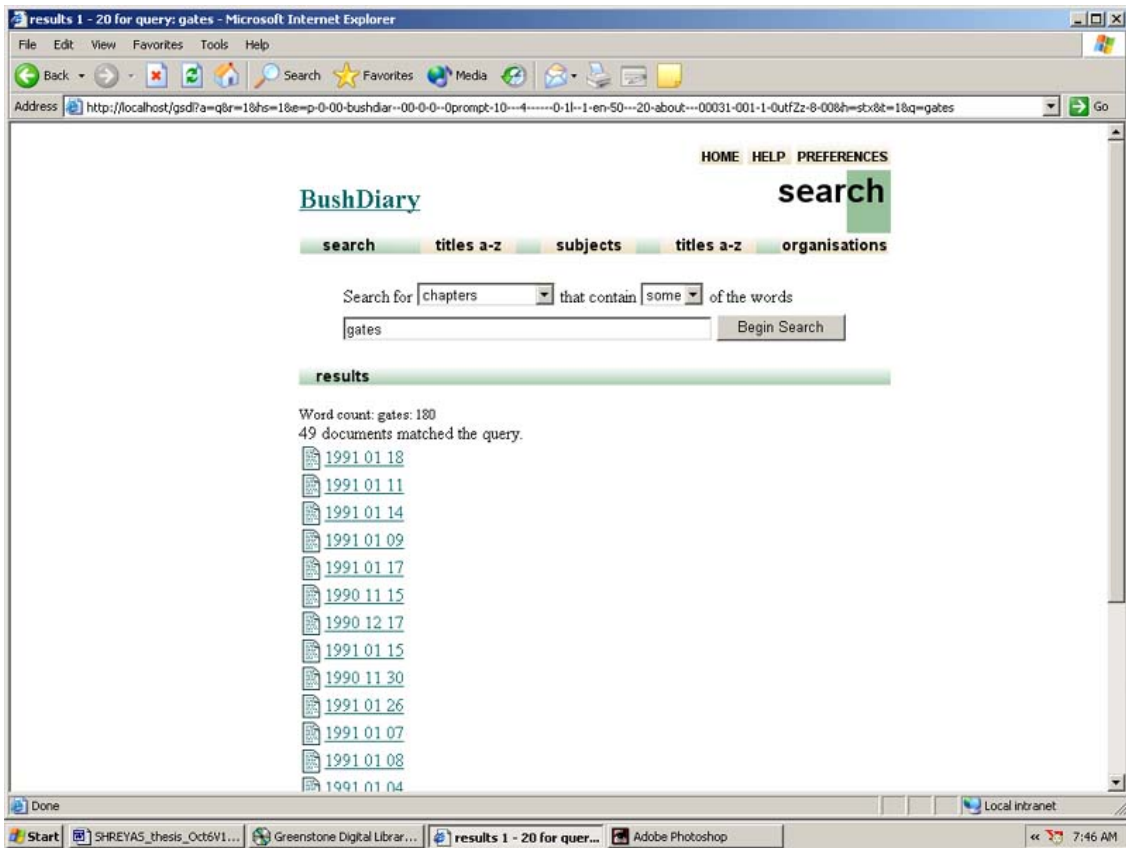


Figure 7: Greenstone Interface - Result Page

VI VISUALIZATION

This project compares user response to timeline-based visualizations of digital library contents to the basic Greenstone-based digital library interface. To perform this comparison, a tool called DigiLine was designed that enables the users to locate and compare data on a timeline. The focus of DigiLine is to show the pattern of occurrence of people and places with respect to events. An important aspect of the tool is that it enables comparisons between two search terms on a timeline. The tool provides interfaces for locating and comparing occurrences of people (or other entities, as input by the user). To judge the relative effectiveness of DigiLine, the same corpus was made available through a prototype of DigiLine and the customized Greenstone interface.

SYSTEM DESIGN

The system design of DigiLine follows the Model-View-Controller (MVC) design pattern. The View consists of the screens used to present the data and its visualization; the Controller consists of the interface for manipulating the analysis engine (used to search and browse the data) and the visualization engine, and the Model is the Greenstone repository. A conceptual view of the system is outlined in Figure 8.

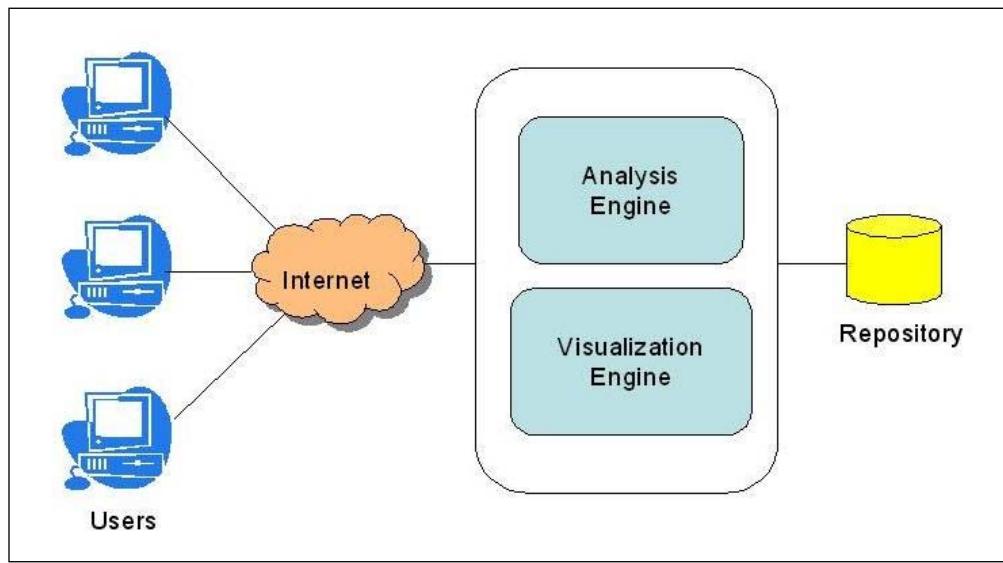


Figure 8: System Overview

TOOLS AND TECHNOLOGY

Regardless of the interface being used to access the library materials, it was necessary to scan and store the documents with proper indexing. As already described, this project makes use of Omni Page Pro OCR Tool to digitize the scanned documents and Greenstone to index and retrieve content from the library. J2EE and the Greenstone Java APIs were used to develop the timeline-based user interface. Net Beans IDE was used for the integrated development. The product can be viewed in Internet Explorer or other standard web browsers.

DIGILINE

DigiLine was designed with the following goals: timeline-based display of results, ability to compare timelines of two sets of search results, fast and easy navigation to located results, and the ability to display the original scanned item and timelines simultaneously. DigiLine internally used Greenstone APIs to connect to the repository. Before developing the interface, feedback and suggestions were solicited from domain experts that included Professors of Political Science.

Time	Date	Event	Description
3:45	4:05		Vice President Quayle
3:45	4:05		Secretary Cheney
3:45	4:05		Gen. Colin L. Powell, Chairman, Joint Chiefs of Staff (JCS)
3:45	4:05		Mr. Scowcroft
3:45	4:05		Mr. Gates
3:58	4:14		Mr. Sununu
4:46		P	The President telephoned Prescott Bush. The call was not completed.
5:13	5:14	P	The President talked with the First Lady.
5:33		P	The President telephoned former President, Ronald W. Reagan. The call was not completed.
5:41		P	The President telephoned former President, Gerald R. Ford . The call was not completed.
5:47	5:51	P	The President talked with Mr. Reagan.
5:58	6:01	P	The President talked on a conference call with: Corazon C. Aquino, President of the Republic of the Philippines

Figure 9: DigiLine Search Interface on Left with Scanned Document on Right.

Figure 9 shows a typical DigiLine screen. The user can enter a keyword and select from a list of predefined events of interest in the left pane. The scanned image of a page in the Presidential Diary collection is shown in the right pane. On clicking submit, the resulting dates that include the keyword are displayed in the right pane (as shown in Figure 10).

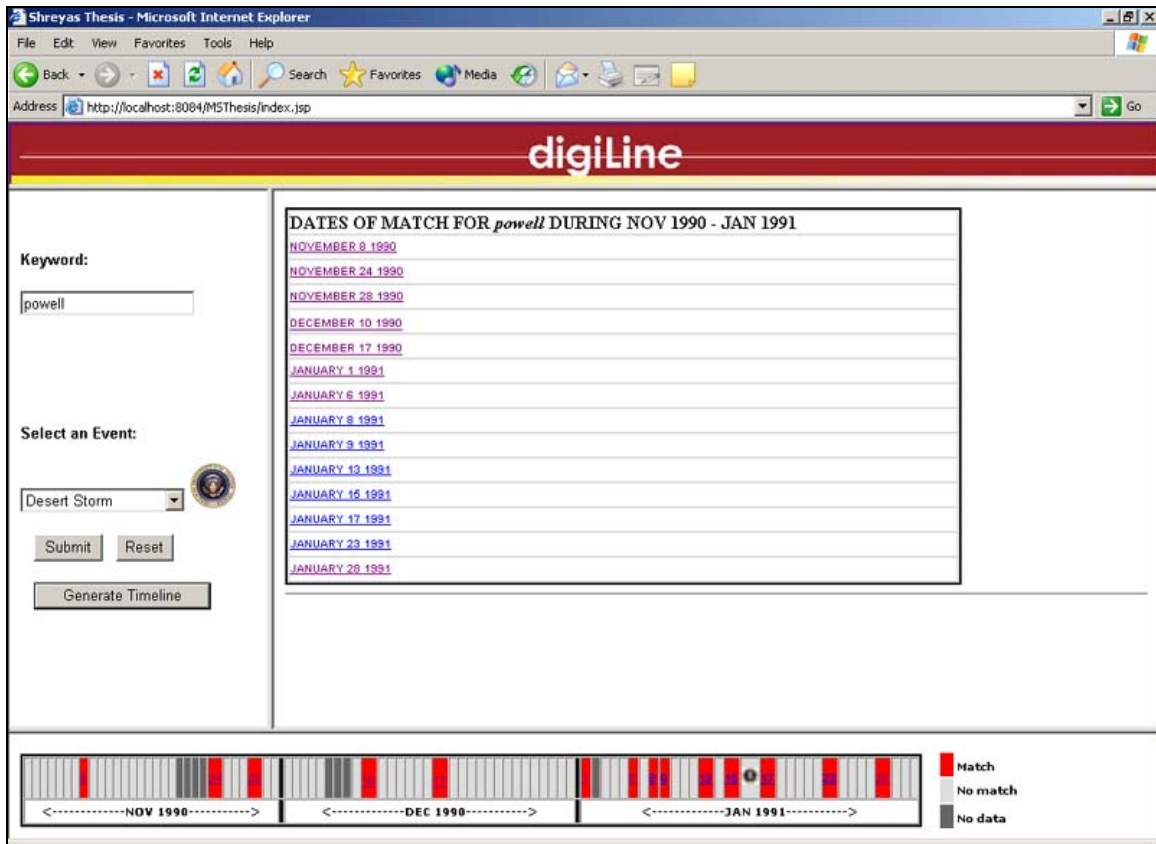


Figure 10: digiLine Results are Shown on Right with Timeline Below.

The “Generate Timeline” button creates a graphical timeline in the bottom pane – the dates that include documents matching the search term are colored. The time line shows data for three months leading to the event selected. At the bottom of the timeline, the month and year are indicated. Of interest to political scientists are how certain events impact the President’s schedule – who is he talking to and how often. The date of the selected predefined event, in this case the beginning of the Gulf War, is identified with a presidential seal in the timeline.

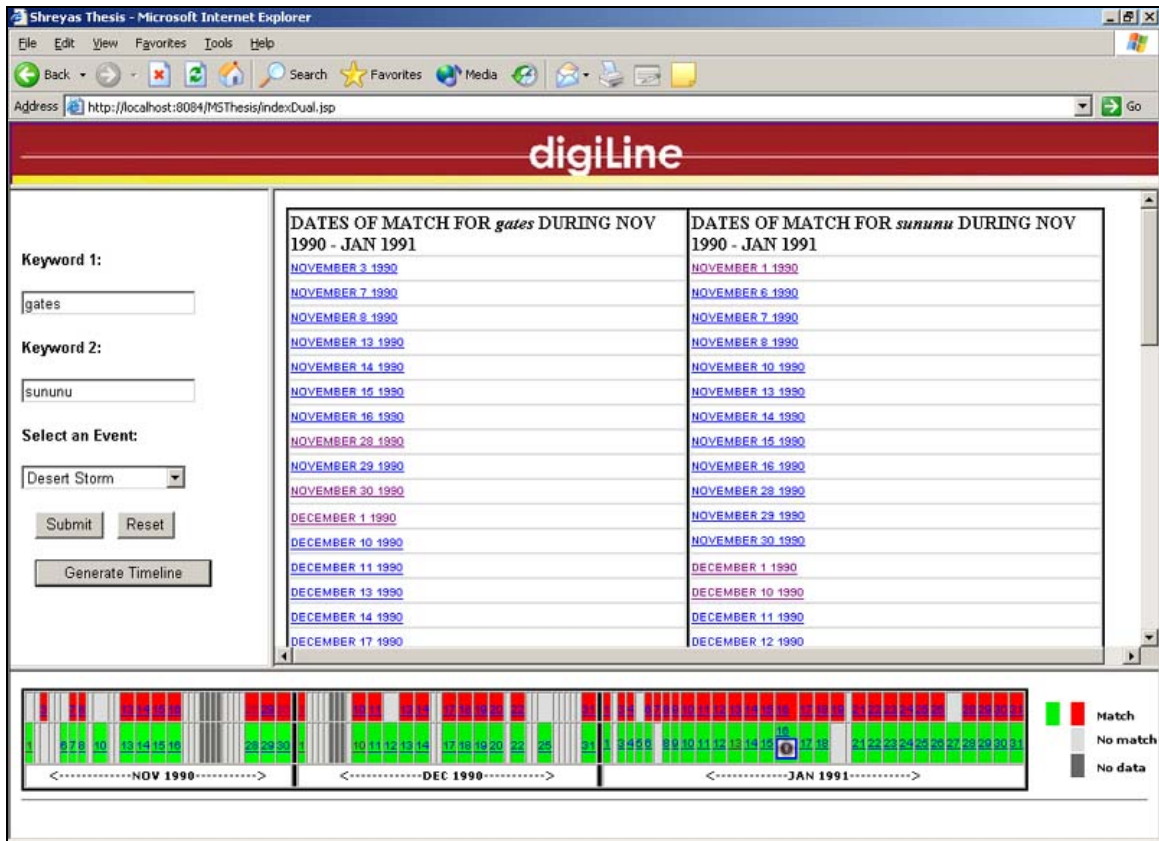
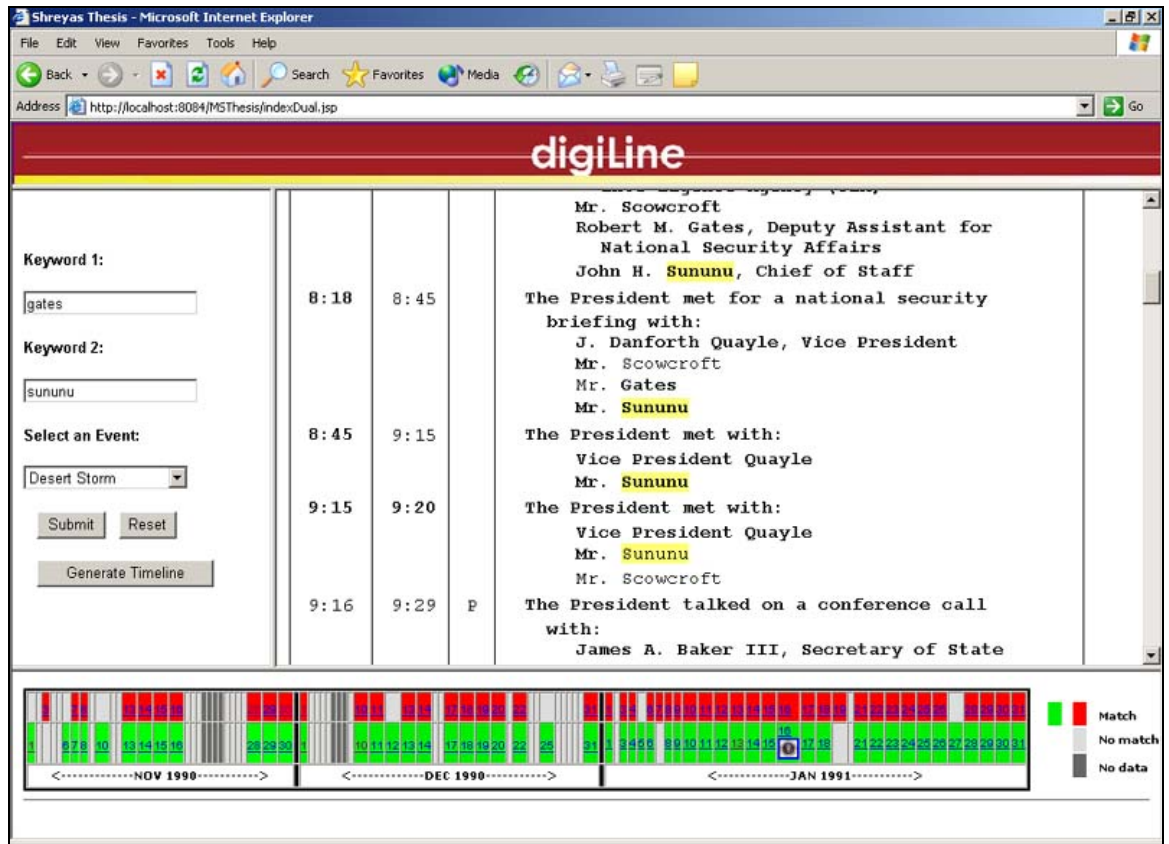


Figure 11: DigiLine with a Timeline Displaying Two Different Search Results

Comparison of how often the President met with key advisors is also of interest. DigiLine users can specify two search terms in addition to selecting an event. Figure 11 shows the comparative timeline in the bottom pane. The occurrence of keyword 1, "powell", is shown in red and that of keyword 2, "kemp", is shown in green. A light grey color indicates no match, and a dark grey color indicates that for that day, there are no records in the diary. If there is a match (red or green)

then if one clicks that cell, the actual diary page for that date is displayed at the top, as shown in Figure 12 with the keywords highlighted in yellow.



The screenshot shows the digiLine web application interface. On the left, there is a search form with the following fields and controls:

- Keyword 1:** Input field containing "gates".
- Keyword 2:** Input field containing "sununu".
- Select an Event:** A dropdown menu currently showing "Desert Storm".
- Buttons: "Submit", "Reset", and "Generate Timeline".

The main content area displays a list of events with the following details:

Time	Event Description
8:18 - 8:45	Mr. Scowcroft Robert M. Gates, Deputy Assistant for National Security Affairs John H. Sununu, Chief of Staff The President met for a national security briefing with: J. Danforth Quayle, Vice President Mr. Scowcroft Mr. Gates Mr. Sununu
8:45 - 9:15	The President met with: Vice President Quayle Mr. Sununu
9:15 - 9:20	The President met with: Vice President Quayle Mr. Sununu
9:16 - 9:29 P	Mr. Scowcroft The President talked on a conference call with: James A. Baker III, Secretary of State

At the bottom of the page, there is a timeline visualization showing a grid of days across three months: NOV 1990, DEC 1990, and JAN 1991. A legend indicates that red bars represent a "Match", green bars represent "No match", and grey bars represent "No data".

Figure 12: DigiLine after Navigation to Document from Timeline

VII EVALUATION

After the initial corpus was digitized and the two interfaces were developed, potential users were selected to gain experience with and suggestions for the Presidential Diary collection and interfaces. The potential users included the students of the Political Science department at Texas A&M University. Many of the respondents were from the Bush School of Government at Texas A&M University. Some of the staff members of the George Bush Presidential Library also responded to the request to participate in the survey. The privacy of the respondents was protected and no personal data was collected. The respondents could opt out of the survey at any point. All respondents were at least 18 years of age. The surveys were conducted on the Bush School premises and other Computer Labs on Texas A&M University campus. No monetary compensation or academic credit was provided for the respondents and they were clearly made aware about this.

The users were asked to use the two interfaces for about 10 minutes each and then asked to fill out a questionnaire. Subjects were randomly assigned into two groups of equal size. One group used the Greenstone-based interface first and the timeline-based interface second, while the other group used the timeline-based interface first and the Greenstone-based interface second.

RESULTS

A total of 12 respondents participated in the evaluation. The user group consisted of both male and female participants. All users had some affiliation with either Texas A&M University or George Bush Presidential Library at College Station. The respondents were a good mix of people from political science and other backgrounds.

The users were divided into two groups - Alpha and Beta. Users of one set (Alpha) were shown the DigiLine interface first, while users the other set were shown the Greenstone interface first. This was done to eliminate bias due to the order in which the interfaces are used.

Table 2: Survey Results

Questions	User group Alpha (Those who used DigiLine first)	User Group Beta (Those who used Greenstone based interface first)
Would you prefer to use the Bush Diaries online, if available, or would you prefer to visit the library and request for the paper version?	Online 100% Paper 0%	Online 83% Paper 17%
Which interface would you like to use for a class project or research?	DigiLine 83% Greenstone 17%	DigiLine 100% Greenstone 0%
Which interface helps you compare the occurrences of people or places?	DigiLine 100% Greenstone 0%	DigiLine 100% Greenstone 0%

Table 2 shows the results from the first three questions. The survey results from both groups show similar trends. Eleven of the twelve users (92%) said that they would prefer to use an on-line interface over having to visit the library and request the paper documents. Eleven of twelve also replied that they would prefer the DigiLine interface over the Greenstone-based interface for a class project or research. All twelve subjects agreed that the DigiLine interface helped compare the occurrences of people or places in the Presidential Library.

In evaluating the DigiLine interface, users commented that the timeline feature is very helpful, but it can be made easier to use and understand. When asked about the strengths of DigiLine, most users responded saying that they were the timeline and better visualization. Among the areas of improvements, some users suggested a more colorful interface. One user suggested the interface should be intelligent and adaptive.

When asked about the strengths of the Greenstone interface, users responded that it is easy to use but has limited features. One user felt it is easy to look for keywords and it is easy to understand the navigation.

When asked about time taken to accomplish a particular search, some users felt that DigiLine comparatively takes more time than the Greenstone-based

interface, primarily because they have to spend more time in understanding the functionality.

When asked “What in your opinion could be a use of this project?”, user responses included applications in homeland security, pattern analysis, teaching, and political science research.

VIII DISCUSSION

All 12 users were able to navigate through both systems with minimal training. Most users were very excited about the utility of the DigiLine tool for their course assignments and research. Almost all users felt that the domain selected for the thesis is very interesting. A high percentage of users commented that a timeline-based presentation of such a digital library would help them understand its contents. One of the users commented that government institutions that make timeline-based data publicly available should use this type of tool.

The evaluation results provide important suggestions on how timeline based digital libraries can be improved and be made friendlier to the users. These suggestions included internationalization support, adaptive user interfaces and extensive help documentation.

As indicated by the variety of applications of the DigiLine interface and the Presidential Diary collection, timeline visualizations of search results can be applied to many tasks. Besides the research activities originally envisioned, applications in the education and security seem promising.

IX CONCLUSION

This thesis explores interfaces for accessing fine-grained time-based digital library content. A characteristic of such content is that the individual entries or documents are not as valuable as the relationships between these entries.

Part of this thesis project involved the digitization of a subset of George Bush's Presidential Diary. This activity identified problems associated with OCR software's application to tabular content. A corpus consisting of the 500 pages corresponding to November 1990 to January 1991 was created.

Two interfaces for searching and browsing this Presidential Diary collection were developed. One was a standard digital library interface created using the customization features included in the Greenstone Digital Library tools. The second interface, DigiLine, provides a timeline-based visualization of search results enabling users to compare the occurrence of different activities and people in the diary.

The evaluation looked at the relative strengths and weaknesses of the two interfaces, and led to some suggestions for the future work in this area. The respondents overwhelmingly indicated their preference for an interface with timeline-based visualizations.

Suggestions for improvements to the DigiLine interface included the addition of tool tips associated with each document, multi-language support, and more general help facilities. The respondents suggested very interesting applications for the tool. The evaluation results indicate that students would use the tool for their class projects.

Timeline visualizations can help users understand time-based and frequency relationships among digital library entries, aiding their analysis of library content. While the Presidential diary collection is atypical, there are plenty of similar collections that either already exist or will be created, especially given the increasing ability to record logs of activity from on-line calendars or other software.

REFERENCES

BROWNSTONE, D. 1994. *Timelines of War: A chronology of Warfare from 100,000 BC to the Present*. Little Brown Publishing, Boston.

CHIEU, H., AND LEE, Y. K. 2004. Query based event extraction along a timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

FEKETE, J., AND DUFOURNAUD, N. 2000. COMPUS: Visualization and analysis of structured documents for understanding social life in 16th century. In *Proceedings of the fifth ACM conference on Digital libraries*.

FLEURIOT, C., MEECH, J., AND THOMAS, P. 1998. Diaries as family communication tools. *Conference on Human Factors in Computing Systems CHI 98 Conference Summary on Human Factors in Computing Systems, Los Angeles, California, United States*, 361 – 362.

GOH, D., AND LEGGETT, J. J. 2000. Patron-augmented digital libraries. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, 153-163.

HORNBAEK, K. 2004. Two psychology-based usability inspection techniques studied in a diary experiment. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction, Tampere, Finland*, 3 – 12.

HSIEH, H., AND SHIPMAN, F. 2002. Manipulating structured information in a visual workspace. In *Proceedings of ACM Conference on User Interface Software and Technology 2002*, 217-226.

HUNTER, J., AND CHOUDHURY, S. 2005. PANIC – An integrated approach to the preservation of complex digital objects using semantic web services, *International Journal on Digital Libraries: Special Issue on Complex Digital Objects*.

KARAM, G.M. 1994. Visualization using timelines. In *Proceedings of Intl. Symposium on Software Testing and Analysis (ISSTA), 1994. In SIGSOFT, ACM Software Engineering Notes. 1994*.

KOVALAINEN M., ROBINSON M. AND AURAMÄKI E. 1998. Diaries at work. In *Proceedings of the 1998 ACM conference on Computer Supported Cooperative Work Seattle, Washington, United States*, 49 – 58.

KUMAR V., FURUTA, R., AND ALLEN, R. 1998. Metadata visualization for digital libraries: interactive timeline editing and review. In *Proceedings of the DL 98 Conference*, 126-133.

MANN, WILLIAM A. 1993. Landscape architecture: an illustrated history in timelines, site plans and biography. John Wiley Publishers, New York.

NARDI, B., SCHIANO, D., GUMBRECHT, M. AND SWARTZ, L. 2004. Why we blog. In *Communications of the ACM Special Issue*, 41 – 46.

THE NATIONAL ARCHIVES SEARCH THE ARCHIVES PAGE. Accessed on April 15, 2005. Author/ creator The National Archives of UK, Web site: <http://www.nationalarchives.gov.uk/searchthearchives/>.

POLITICAL SCIENCE RESOURCES, Western Connecticut State University accessed on July 17, 2005, Author/ creator: Western Connecticut State University, Web site: <http://www.wcsu.edu/socialsci/polscres.html>.

SHIPMAN, F., HSIEH, H. AND AIRHART, R. 2001. Analytic workspaces: Supporting the emergence of interpretation in the Visual Knowledge Builder. In *Proceedings of Interact 2001, Tokyo, Japan*, 132--139.

SMITHSONIAN INSTITUTE. 1993. Smithsonian timelines of the ancient world. Smithsonian Institution Press, Washington D.C.

TAKAHASHI, J., LOVERANCE, R., KUSHIDA, T., HONG, J., SUGITA, S, KURITA, Y., RIEGER, R., MARTIN, W., GAY, G., AND REEVE, J. 1998. Global digital museum: multimedia information access and creation on the internet. In *Proceedings of the third ACM Conference on Digital Libraries Table of Contents Pittsburgh, Pennsylvania, United States*.

THE THOMAS ONLINE DATABASE. Congressional Record of the 109th Congress (2005-2006) Accessed on July 16, 2005, Author/ creator: The Thomas Database. Web site: <http://thomas.loc.gov/home/r109query.html>.

WALDMAN, C. 1994. Timeline of Native American history. Prentice Hall, New York.

APPENDIX A
AN ACTUAL PAGE OF THE DIARY

THE WHITE HOUSE			THE DAILY DIARY OF PRESIDENT GEORGE BUSH	Page 3
LOCATION			DATE	JANUARY 16, 1991
THE WHITE HOUSE WASHINGTON, D.C.			TIME	12:05 p.m. ^{DAY} WEDNESDAY
IN	OUT	PHONE	ACTIVITY	
12:06	12:12	P	The President talked with his son, Neil M. Bush.	
12:43	12:44	P	The President talked with his Deputy Assistant, Patricia A. Presock.	
12:56			The President returned to the Oval Office. He was accompanied by Secretary Baker.	
12:56	1:50		The President met with: Secretary Baker	
12:58	1:46		Mr. Scowcroft	
1:00	1:45		Mr. Sununu	
1:09		P	The President telephoned former President, Ronald W. Reagan. The call was not completed.	
1:12	1:17	P	The President talked with former President, Gerald R. Ford.	
1:22	1:24	P	The President talked with Mr. Reagan.	
1:50	2:06		The President met with Mr. McGroarty.	
1:56	1:58	P	The President talked with Senator Domenici.	
2:01	2:03	P	The President talked with the First Lady.	
2:06			The President went to the Cabinet Room.	
2:06	2:59		The President participated in a meeting with members of the President's Education Policy Advisory Committee. For a list of attendees, see <u>APPENDIX "A."</u> Members of Press in/out	
2:59			The President returned to the Oval Office.	
2:59	3:02		The President met with: Mr. Sununu	
2:59	3:13		Mr. Scowcroft	
3:13	3:18		The President met with Mr. Sununu.	

APPENDIX B

QUESTIONNAIRE

Student/ Non student _____

Major (if student) _____ (OPTIONAL)

Given Computer Interfaces A and B:

1. Which interface you would like to use for a class project or research? A B
2. What is the feature that you like most about the interface A?
3. What is that you like least about the interface A?
4. What is that you like most about the interface B?
5. What in your opinion could be a use of this project?
6. Would you prefer to use the Bush Diaries online, if available, or would you prefer to visit the library and request for the paper version?
7. Which interface helps you compare the occurrences of people or places?
8. To find out a particular piece of information, how much time did it take you for interface A and interface B?
9. Your comments for the improvement of interface A
10. Your comments for the improvement of interface B

APPENDIX C

INFORMATION SHEET

PATTERNS IN PRESIDENTIAL DIARY OF 41st PRESIDENT GEORGE BUSH

You have been asked to participate in a research study about digital libraries, specifically a digital library about PATTERNS IN THE PRESIDENTIAL DIARY OF 41st PRESIDENT GEORGE BUSH. You selected to be a participant because of your response to the notice put up by the researcher. A total of 30 people have been asked to participate in this study.

The purpose of this study is to determine what methodology can best help a user find out useful information from a timeline based digital library.

If you agree to be in this study, you will be asked to compare two computer based user interfaces and fill out a questionnaire about their relative strengths.

This study will only take about 30 minutes. The risks associated with this study are Minimal (as associated with less than 30 minutes of Computer Use). The benefits of participation are none.

This study is anonymous and your name or any personal identification will NOT be collected. It is optional for students to report major. Reporting major is relevant to the research.

Your decision whether or not to participate, will not affect your current or future relations with Texas A&M University or any of its departments. If you decide to participate, you are free to refuse to answer any of the questions that may make you uncomfortable. You can withdraw at any time.

This research study has been reviewed by the Institutional Review Board- Human Subjects in Research, Texas A&M University. For research-related problems or questions regarding subjects' rights, I can contact the institutional Review Board through Ms. Angelia Raines, Director of Research Compliance, Office of Vice President for Research at (979) 458-4067.

You can contact the following with any questions about this study.

Shreyas Kumar at kshreyas@hotmail.com or at (979) 846 3686

Dr. Frank M. Shipman III at shipman@csdl.tamu.edu or at (979) 862 3216

Thank you for your time.

Sincerely,

Shreyas Kumar

Principal Investigator

VITA

Name: Shreyas Kumar

Address: c/o Dr. Frank Shipman

Department of Computer Science

301 Harvey R. Bright Bldg,

College Station, TX 77843-3112 USA

Education: 2005, M.S. Computer Science, Texas A&M University

1999, B. Arch., Indian Institute of Technology, Roorkee

Experience: 2004 -2005, Webmaster, International Center, Texas A&M Univ.

1999-2003, Software Engineer, Perot Systems Corporation