Using Information Visualization to Support Access to Archival Records

Robert B. Allen, PhD

ABSTRACT. As more archival metadata and archived records are available online, providing effective interfaces to those materials is increasingly important to give users effective access. We describe five approaches from the hypertext and visualization research communities which can be used to improve such access. We consider: 1) Navigating hierarchical structures, 2) Illustrating networks of relationships, 3) Viewing processes, 4) Using time and space as organizing structures, and 5) Spatializing arguments and discussions. We address particular challenges in applying these visualization techniques to archival records and consider novel challenges introduced by them for both archival organization and for visualization research.

KEYWORDS. Advanced Interfaces, Argumentation Systems, Browsing, Electronic Records, Human-computer Interaction (HCI), Searching, Interaction, Visualization,

INTRODUCTION

There are many ways that archival materials can be used: Consider the citizen who wants background about the formation of a new national park, or the Congressional aide who is collecting information about the effectiveness of veteran's hospitals. Similar types of needs might be found at the state and local level with the history researcher who is studying the development of water-conservation policies or the newspaper reporter investigating possible corruption in the construction of a highway.

Robert B. Allen, PhD, is Professor of Practice, College of Information Studies (CLIS), at the University of Maryland, Hornbake Hall, Suite 4121E, University of Maryland, College Park, 20742-4345 (E-mail: allen@acm.org). Dr. Allen teaches courses on digital libraries and information management. Dr. Allen is the Chair of the Publications Board of the ACM and was Editor in Chief of the ACM Transactions on Information Systems for 10 years. Before moving to Maryland, he was Senior Scientist in the Information Science Research Group at Bellcore. He received his PhD in Experimental Psychology from the University of California, San Diego (UCSD), and joined Bell Laboratories in 1978.

In these tasks, the researchers would benefit from understanding the relationships among government agencies, sets of documents, and the threads of debates. Current interface tools, such as those supporting the Archives Resource Center (ARC) and Access to Archival Data (AAD) have only HTML-based Web form interfaces. However, it is possible to imagine many more-advanced interfaces. Graphical, visualization, and hypertext-like interfaces have many advantages. The high degree of interactivity available with such interfaces allows rapid exploration, helps improve understanding, and provides Information visualization takes advantage of information structure and therefore is useful for government materials which are generally highly structured records. We may say that the visualization "spatializes" the structure: That is, conceptual structures are laid out on a graphical display. Structure can be derived from many sources including attributes such as metadata descriptions of archival records such as the Encoded Archival Description (EAD). (Library of Congress).

A variety of visualization and hypertext techniques have been developed; we discuss five families of these techniques which we believe are particularly applicable for accessing Federal records and archives. A general overview of visualization techniques can be found in (Card, Mackinlay, and Shneiderman 1999). While visualization is sometimes used for exploring records in data mining and data warehouses (e.g., Berson and Smith 1997) it has not been used much for exploring archived materials.

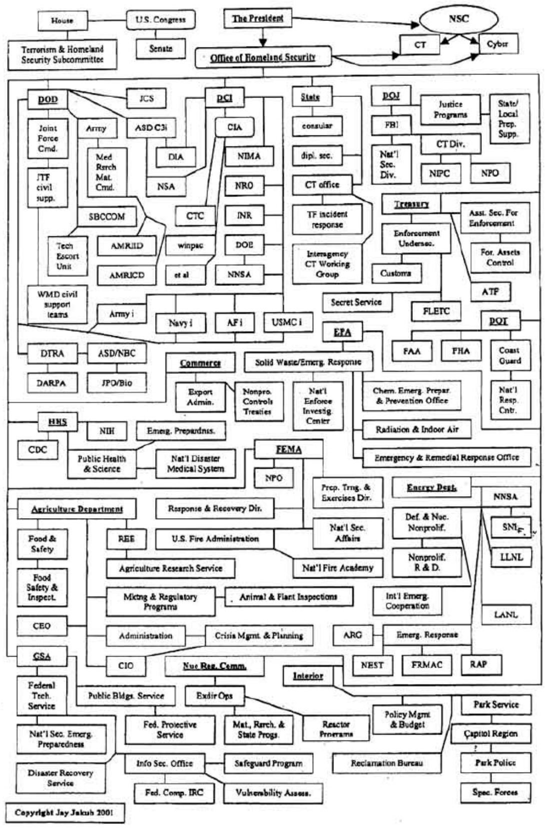
There are, of course, limitations and difficulties in using information visualization interfaces. These interfaces require more development effort than traditional interfaces and even if the interfaces are well designed, the users may require more training than with traditional interfaces because of the richness of the features they offer.

APPROACH I:
NAVIGATING HIERARCHICAL STRUCTURES

Organizations such as the federal and state governments are large and disjoint. It is difficult to keep track of the structure and the information they produce. Better models for combining views of information across government organizations should reduce the tendency to develop silos and stovepipes in which even public information is often effectively accessible only from within its originating organizations.

The structure of the government as defined by statute is generally hierarchical. Figure 1 shows the major organizations within the proposed Department of Homeland Security (from Harriman 2002). While the hierarchical structure should facilitate access, the tangled static graphic makes use difficult. It would probably be much easier if the organization could be examined at different levels of detail using an interactive hierarchical browser. Several hierarchy viewers have been developed (e.g., Egan et al. 1989) (Nation et al. 1997) and they are now familiar in tools such as the Windows file manager. Hierarchical browsing can be particularly helpful but it can have even greater power in conjunction with search. For instance, a search on “Alaska” could give pointers to all agencies dealing with Alaska. Of course, several different strategies could be useful for the indexing of those agencies. To the extent they are available, functional descriptions of each agency and department could provide richer indexing.

FIGURE 1. A chart showing the organizations that were merged into the Department of Homeland Security. While the details are difficult to read in this reproduction, the main point is the complexity for the user to understand the relationship of parts of the organization from type of display.



Not only is the structure of government quite complex; it evolves through time. For instance, the recently proposed Department of Homeland Security includes numerous agencies drawn from other organizations. Making those changes clear for a user would be a challenge. Moreover, in some cases as a result of such changes an agency's records may be spilt between the Library of Congress and the National Archives**.** So, giving the researcher that background may also help the researcher to locate the records. Perhaps we could extend existing visualization techniques with before-and-after shots and a narrative that describes how the transition occurred.

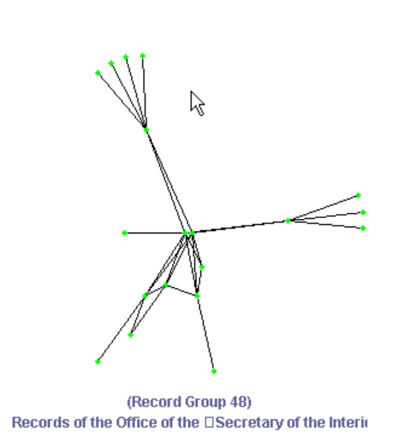
APPROACH II:
MAPS OF RELATIONSHIPS

While hierarchies provide a simple framework that helps to orient the user, other content has a less-regular link structure. For such irregularly structured material, a local view of the links may still be useful. That is, the researcher can view the immediate neighbors, and perhaps, the neighbors of those neighbors.

As an example, archival Record Groups are linked to other Record Groups. Once a researcher finds one promising Record Group that researcher might want to locate other related records. To explore linking based on similarity of records; it is easy to imagine navigation as flying through these Record Groups. A browser allows users to find materials that are similar to those he/she is already viewing. Suppose a user was searching for information about the fishing rights of native Alaskans in rivers. If that person had found the Fish and Wildlife Service RG#022 he/she could browse to other related records. For the description of each Record Group, the *Administrative Guide* (Matchette 1995) includes "Related Records". While there is some debate about the value of Record Groups, we use them here primarily to illustrate the visualization technique.

These views of node neighborhoods may be thought of as "maps" which give an overview of the neighborhood of related concepts. For easy viewing, the nodes may be spread apart across the screen. This spacing may be accomplished with a "self-organizing map" (SOM) (e.g., Lin, Soergel, and Marchionini 1991) which cluster search results into conceptual categories. Specifically, we (Stedfeld and Allen in preparation) used a "mass-spring model" (Provot 1995) to spatialize the relationship among the concepts. Figure 2 shows a prototype interface for the Record Groups related to the Fish and Wildlife Service (RG#022). Three major groups are apparent: In the upper left is the Coast Guard Group (RG#026); on the right are agencies relating to oceanography and water resources; in the lower left is the Department of the Interior (RG#048).

FIGURE 2. Record groups related to the Fish and Wildlife Service. The labels describing the Record Groups are presented on mouse rollover.



The map in Figure 2 is generated from a single, uniform measure of relatedness among the Record Groups. Other measures of relatedness could also be explored, such as the distance between items in a hierarchical classification like the GPO Superintendent of Documents Classification (MacGilvray 1993) or the co-occurrence of citations in the literature.

A working prototype interface (illustrated in Figure 2) shows relationships at the level of Record Groups, but the rich detail of SubGroups could also be used. For instance, 22.4.1. has Records of the Division of Alaskan Fisheries; this could be linked to the records of the Alaska District of the Army Corps of Engineers (RG#077.10.1) or the Records of the Alaska Division Bureau of Native American Affairs (RG#075.11). For defining such detailed connections, it may be useful for a human expert to specify additional links.

APPROACH III: VIEWING ORGANIZATIONAL PROCESSES

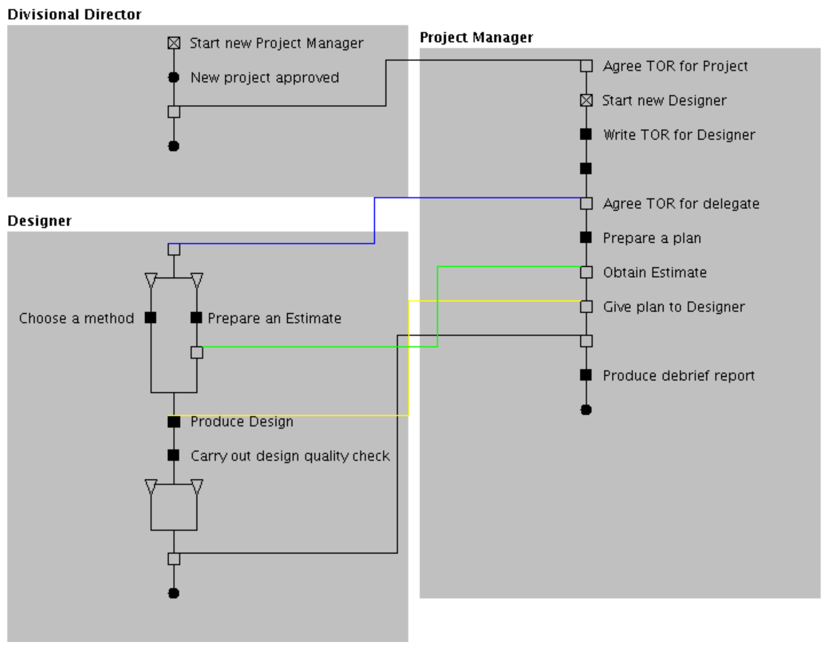
Visualization can be used to view not only structure but also processes. Government includes many activities which have regular decision points. There are many examples such as steps in the procurement processes, steps in a judicial appointment, or stages in the review of documents for external release. In some cases, the process may be the main point. In other cases, it can provide a structure for understanding a collection of inter-related documents.

There is, currently, a great deal of work on developing wrappers for preservation of digital resources. For instance, the Metadata Encoding and Transmission Standard (METS) (Beaubien) approximately implements the Open Archival Information System (OAIS) Reference Model (CCSDS 2002). One component of METS is descriptive metadata. Some of our recent work has developed descriptive metadata for project management data from NASA Goddard Space Flight Center (Allen and Templeton in preparation) (Hodge, Allen, and Templeton 2003). For instance, this could be used to describe the process by which an agency approves a new initiative. Specifically, we employed a modeling technique known as Role Activity Diagrams (RADs) (Ould 1995). This models organizational roles and interactions between those roles. The RADs are metadata which can explain the organizational activities that led to the creation of a document. As shown in Figure 3, we developed XML specifications for RADs and generated a graphical Java display from it (Figure 4). The graphical display can be used describe where in the organizational workflow records originated. A similar approach could be applied to keeping track of versions of legislation in the U.S. Congress as it passed through various committees and amendments (Dove 1997) as well as business documents (Zisman 1997).

FIGURE 3. Fragment of XML coding for a Role Activity Diagram (RAD). Note the definition of specific Roles and of Interaction Points between the Roles.

```
<?xml version="1.0" encoding="UTF-8"?>
<RoleActivityDiagram
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="RADsch111.xsd">
  <RoleSec>
    <Role ID="R1" type="Project Manager">
      <StartRole ID="R1.SR" description="Start new Project Manager" from="R1.EE1" to="R1.IP1"/>
      <ExternalEvent ID="R1.EE1'"description="New project approved" to="R1.SR"/>
      <InteractionPoint ID="1.IP1" from="R1.SR">
        <Actor identification="000-00-0000"/>
      </InteractionPoint>
    </Role>
    <Role ID="R.2" type="Designer">
     <Activity ID="R2.A1" description="Choose a method" from="R2.M1" to="R2.M2">
       <Actor identification="000-00-0001"/>
     </Activity>
     <Activity ID="R2.A2" description="Prepare an Estimate" from="R2.M1" to="R2.A3">
       <Actor identification="000-00-0002"/>
     </Activity>
...
```

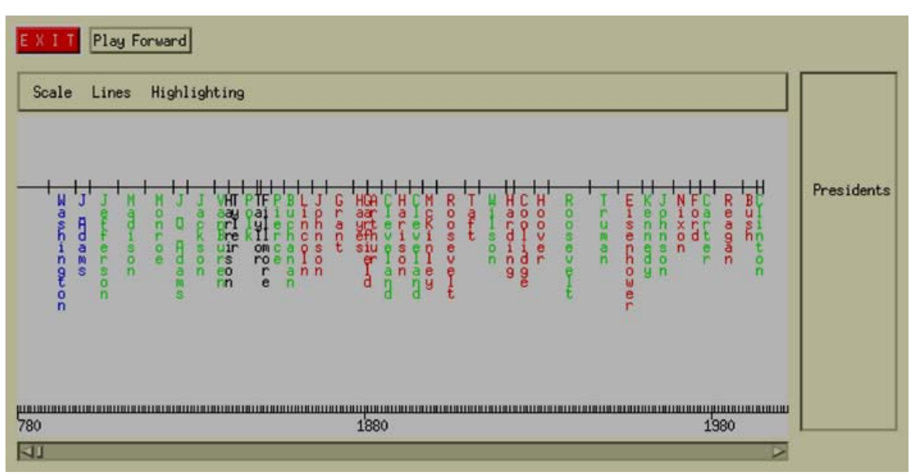
FIGURE 4. In our interface for viewing the organizational content in which records are created, Role Activity Diagrams (RADs) are reconstructed from an XMLSchema.



APPROACH IV:
SPACE, TIME, CAUSATION, AND NARRATIVE

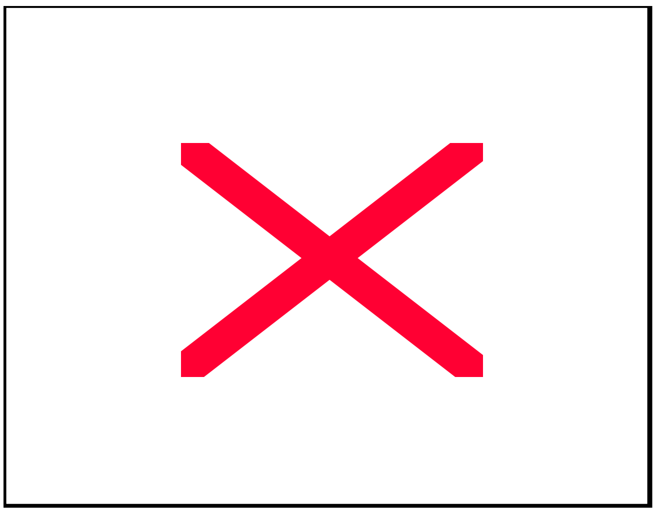
Events occur in time and space. Time is an integral dimension for events and it can be used to structure visualizations in timelines. Figure 5 shows a timeline of the terms of the U.S. presidents (Allen 1995). Much richer detail can be added; the events within a presidency could be organized by extension of such a timeline. An elaboration of such an interface to include important events during an administration might be useful, for instance, for researchers at the Presidential Libraries.

FIGURE 5. Interactive timeline for U.S. presidents. Attributes such as political party can be highlighted.



Time is also useful as an organizational framework. The date of accession for Records or even for Series can be useful for organizing those records. For instance, Figure 6 shows a sketch of a proposed interface for the history of an individual record group and even individual documents and records -- in this case, for the Interstate Commerce Commission. Relevant historic events could be shown on a second set of timelines. Because the documents often discuss the historic events, there should be a linkage from the documents to the historic events. For instance, a record created in 1914 may mention somebody born in 1872 and married in 1898. Overall, this work can be viewed as part of a larger project in the author's laboratory dealing with historical events (Allen 2004). That project attempts to make standard descriptions of those events and tools for interacting with them (e.g., Allen and Schalow 1999).

FIGURE 6. Sketch of a possible timeline-oriented Record-SubGroup viewer. Specifically, this shows the dates associated with the creation of SubGroup records for the Interstate Commerce Commission (RG#034).



The Records could be coded by space as well as time. In other words, a graphical geospatial information system (GIS) could help to orient the users. In this case, the spatial structure could be coded with geo-referenced data metadata. It could also include spatio-temporal views of events described in the document. For instance, military history such as shown in Table 1 (U.S. Army) could be animated with a GIS map. Such interfaces for presentation of space and time are still poorly explored. As with temporal coding, several spatial dimensions of the documents could be employed including spatial information relevant to the document itself such as where it is stored.

TABLE 1. A sample of an official military history, which could be presented with a spatio-temporal interface.

With the hill taken 11 July 1944, the Division went on the defensive until 26 July. Exploiting the St. Lo break-through, the 2d Division advanced across the Vire to take Tinchebray 15 August 1944. The Division then moved west to join the battle for Brest, the heavily defended fortress surrendering.

Beyond simply producing an ordered sequence of events, histories describe the ways in which events cause other events. Allen and Jun (in preparation) have explored causal links for science texts. The plots in children's stories may be modeled as a chain of causal explanation (Allen and Acheson 2000). It is should also be possible to describe such links for causal explanations and historic events. Furthermore, such causal explanations could be the basis for navigating documents.

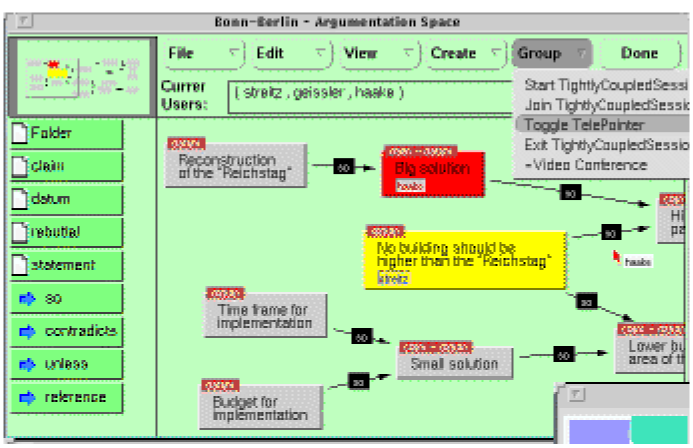
A related model, which is perhaps closer to hypertext than to visualization would create a type of guided tour (e.g., Zellweger 1989) or even a guided multimedia animation. For instance, navigation support for the Nixon tapes could include several predetermined paths through descriptions of the content. Moreover, they could provide limited browsing beyond each of those paths.

APPROACH V: SPATIALIZED ARGUMENTATION

Many activities in organizations such as government are the result of discussion and debate. The hypertext community has developed “argumentation systems” for organizing the issues in a group discussion. For example, Figure 7 shows (hypothetical) arguments and counter-arguments for a decision about moving the German capital from Bonn to Berlin and the related question of whether the Reichstag should be rebuilt (Streitz et al. 1992). Although we won't discuss all the details here, the propositions and issues are shown as rectangular nodes and they are linked by predicated such as “so”, “pros”, “cons”, “conclusions”, and “evidence”.

Argumentation systems could be used for descriptions of many governmental and organizational activities such as hearings and speeches. There are many such discussions in the Federal government such as committee hearings and Congressional debates that lead up to these. These systems have also been used to provide explanations of design decisions (i.e., “design rationale”) (Shum 1996) and they could be applied to many government engineering projects. One concern is that this type of presentation is a non-chronological perspective on arguments which may require interpretation. To avoid bias, standards may need to be established to provide balanced presentations. Alternatively, these tools could be made available to the public to develop their own interpretations of the discussions.

FIGURE 7. An argumentation system spatializes the components of an argument. In this case, a discussion of German government policy.



EXTENSIONS AND CONCLUSIONS

Several types of information visualization tools for improving access to archival material have been explored. These applications are useful extensions of existing technologies but we believe they can be pushed further. For instance, many of the current tools will be challenged by the size and complexity of these data sets and these tasks could provide an opportunity to investigate how to adapt the tools to such large-scale problems. For instance, wall-sized displays (e.g., Tan and Czerwinski 2003) might show details of many government agencies in one screen.

Beyond identifying isolated visualization tools we can consider how they may be combined into a system. Putting several of these visualization and hypertext tools together along with other data management tools we could create an archival-researcher's or a record-manager's workbench for supporting complex searches and combining materials from many sources. Naturally, in developing such a system, a careful evaluation of user needs would be useful along with usability testing (Nielsen 1993). Moreover, these interfaces will be most useful if their development is based the needs of actual users engaged in complex tasks (e.g., Toms and Duff 2002). Finally, the question of coordination across visualization tools is largely unexplored; for instance, we could explore how a timeline could best be incorporated with an argumentation system.

ACKNOWLEDGMENTS

From the College of Information Studies at the University of Maryland, I thank Bruce Dearstyne, Susan Davis, and Tanya Marshall. Other collaborators and students including Manav Kher, Jun Luo, and Clay Templeton have also provided advice. From NARA, I thank Bob Matchette and Bruce Ambacher. I also thank Richard Cox for his comments of a draft. I would like to thank Eric Stedfeld at NYU for extensive discussions about user interfaces for archival materials and for reading an early draft of this paper.

REFERENCES

Allen, R. B., June 2004. "Building an Event Gazetteer." *ACM/IEEE Joint Conference on Digital Libraries*, in press.

Allen, R. B. 1995. "Timelines as Information System Interfaces." *Proceedings International Symposium on Digital Libraries*. 175-180.

Allen, R. B., and J. A. Acheson. 2000. "Browsing Structured Multimedia Stories." *Proceedings ACM Digital Libraries Conference*. 11-18.

Allen, R. B., and Jun L. in preparation. "Interactive Cognitive Organizers: Models of Causation in Science Texts."

Allen, R. B., and J. Schalow. 1999. “Metadata and Data Structures for the Historical Newspaper Digital Library Project.” *Proceedings ACM Conference on Information and Knowledge Management*. 147-153.

Allen, R. B., and C. Templeton. in preparation. “Organizational Context for Digital Preservation of Project Documentation.”

Beaubien, R. METS: An Overview and Tutorial. http://www.loc.gov/standards/mets/-METSOverview.html.

Berson, A., and S. J. Smith. 1997. *Data Warehousing, Data Mining, and OLAP*. New York: McGraw-Hill.

Card, S., B., J. Mackinlay, and B. Shneiderman. 1999. *Information Visualization: Using Vision to Think*. San Francisco: Morgan-Kaufmann.

CCSDS. 2002. Reference Model for an Open Archival Information System (OAIS). ISO 14721:2002.

Dove, R. Feb. 1997. Enactment of a Law (Senate). http://thomas.loc.gov/home/enactment/-enactlawtoc.html.

Egan, D., E., J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum. 1989. “Formative Design Evaluation of SuperBook.” *ACM Transactions on Information Systems*. 7 (1): 30-57.

Harman, J. Winter 2002. Department of Homeland Security. From the Web site of Congresswoman Jane Harman, http://www.house.gov/harman/terrorism/homeland.html.

Hodge, G., R. B. Allen, and C. Templeton. 2003. “Project Metadata: Development of the Goddard Core.” *Issues in Science and Technology Librarianship*

Library of Congress. “Encoded Archival Description (EAD).” ://www.loc.gov.ead/.

Lin, X., D. Soergel, and G. Marchionini. 1991. “A Self-organizing Semantic Map for Information Retrieval.” *Proceedings of the ACM SIGIR Conference*. 262-269.

MacGilvray, M. W. 1993. *GPO Classification Manual: A Practical Guide to the Superintendent of Documents Classification System*. Washington, DC.

Matchette, R. B. 1995. *Guide to Federal Records in the National Archives of the United States*. National Archives and Records Administration. Also: http://www.archives.gov/-research_room/federal_records_guide/

Nation, D. A., C. Plaisant, G. Marchionini, and A. Komlodi. 1997. “Visualizing Websites Using a Hierarchical Table of Contents Browser.” *Human Factors and the Web*.

Nielsen, J. 1993. *Usability Engineering*. San Diego: Academic Press (Professional).

Ould, M. A. 1995. *Business Processes: Modelling and Analysis for Re-Engineering and Improvement*. John Wiley and Son, Ltd.

Provot, X. 1995 “Deformation Constraints in a Mass-Spring Model to Describe Rigid Cloth Behaviour.” *Graphics Interface*. 147-154.

Shum, S. B. 1996. “Design Argumentation as Design Rationale.” In *The Encyclopedia of Computer Science and Technology*, edited by A. Kent and J. G. Williams, 95-138. New York: Marcel Dekker.

Stedfeld, E., and R. B. Allen. in preparation. “Supporting Browsing of Archival Record Groups.”

Streitz, N., J. Haake, J. Hannemann, A. Lemke, W. Schuler, H. Schutt, and M. Thuring. 1992. “SEPIA: A Cooperative Hypermedia Authoring Environment.” *Proceedings ACM Hypertext*. 11-22.

Tan, D., and M. Czerwinski. 2003. “Information Voyeurism: Social Impact of Physically Large Displays on Information Privacy.” *Proceedings ACM SIGCHI: Human Factors in Computing Systems*.

Toms, E. G., and W. Duff. 2002. “I spent 1½ hours sifting through one large box….: Diaries as information behavior of the archives user: Lessons learned.” *Journal of the American Society for Information Science and Technology*. 53:1232-1238.

U.S. Army. U.S. Army Center of Military History. ://www.army.mil/cmh-pg/.

Zellweger, P. T. 1989. “Scripted Documents: A Hypermedia Path Mechanism.” *Proceedings ACM Hypertext*. 1-14.

Zisman, M. D. 1977. “Representation, Specification and Automation of Office Procedures.” Ph.D. diss., Wharton School, University of Pennsylvania.