"The Library of Congress at a Glance": Text Visualization and Reference Rooms Without Walls

by Lee A. Gladwin[1]
Center for Electronic Records,
National Archives and Records Administration

You should "be able to see the Library of Congress at a glance" declared Ben Shneiderman, Head of the Human-Computer Interaction Laboratory at the University of Maryland. He was one of many presenters at the Advanced Information Processing & Analysis Symposium which was held at Tyson's Corner, Virginia between March 22nd and March 24th, 1994. The symposium was organized by the Advanced Information Processing & Analysis Steering Group which represents the intelligence community. A primary goal of this organization is to provide liaison between intelligence analysts and potential contractors. Analysts are responsible for assimilating and synthesizing information from a variety of sources and producing digests of the request in timely fashion. In effect, they face the same overwhelming flood of information that all researchers do. They need some technology which allows them to visualize the search environment at a glance, filter out irrelevant information and focus in on what is critical to their tasks. Text visualization is one of the technologies being explored.

Shneiderman provided an example of how the University of Maryland Library holdings could be represented by their Macintosh-based TreeViz program as a series of 100 boxes on a computer screen. Each box represented a Dewey decimal-based classification, such as History, Science or Philosophy. The size of the box was proportional to the size of the particular holdings. Boxes were color coded to indicate rate of use. Boxes were arranged in hierarchies, so that one could descend from a general box, such as History, to a more detailed box-representation of books in various fields of history. Not all intuitive user interface designers agree that more is better. The screen design suggested by Shneiderman could be overwhelming for some users. Other designs and retrieval technologies are examined below.

Text Retrieval and Visualization Systems

Since the majority of information requested by users is textual in nature, a great deal of research is being done to find improved ways of clarifying the researcher's information needs (queries) and comparing them with text content in the database. The effectiveness of any text retrieval system is the degree to which it satisfies "an information need" [Croft, 9]. Does it retrieve most of the relevant documents from the prospective document pool? To satisfy these criteria, any system must be capable of "representing a user's information problem or need, representing the content of text documents, and comparing these representations to decide which documents should be retrieved" [Croft, 9]. There are two approaches: statistical and knowledge-based. The former is based upon the notion that words occurring less frequently in documents are more important than those frequently found. Knowledge-based approaches are more concerned with the human aspects of information retrieval [Croft, 10]. These text retrieval approaches are fundamental to text visualization.

Before a document collection can be visualized, each document must be retrievable. To be retrievable, each must be represented within the retrieval system. Conventionally, documents are represented in an inverted index file. Each keyword is individually indexed and cross-referenced with the documents containing it. An inverted file is the "set of indexes for all allowable terms and attribute values" [Salton, 232]. This file might be imagined as a spreadsheet with column headers consisting of document identifiers and intersecting rows of keywords. At each column-row intersect would be a "O" if the terms does not appear in the document or a "1" if it does. A further refinement is the addition of term locations to the index giving the frequency of keyword occurrence in a document, and the precise location of the term; e.g., Document 35O, paragraph 11, sentence 3, word 7. Since not all terms are created equal, keywords are weighted by their importance within the document. Frequently occurring words such as "the" or "and" receive lesser weights than infrequently occurring keywords such as nouns. Keyword weights, therefore, are also added to the index [Salton, 231 - 239]. Indexing may be done manually by human experts or automatically by a program such as those described below.

When indexing is done automatically, keywords are identified, their weights computed and combined to produce a document vector that may be compared with other document vectors to compute the degree of similarity and provide a

basis for graphical mapping of document relationships [Salton, 275 - 290, 304 - 308]. Following the description of both documents and queries in terms of their vectors, retrieval is executed on the basis of a "computation of query-record similarities" [Salton, 275].

TASC's TEXTVIZ program follows the knowledge-based approach. Multiple documents are first scanned via an optical Character Recognition (OCR) device into the database from which "features" (symbols, keywords, phrases such as "leaders, "house", "Bogota") are then extracted using a natural language processor. Features are then converted to "feature vectors" or unique identification codes which allow the system to compute "the degree to which a specific concept is correlated with a document" [Textviz, 2]. Text features or concepts may then be "mapped to points in a graphical space (text map)" where documents dealing with similar concepts are clustered together. Dissimilar documents are spaced farther apart on the screen. In TASC's TEXTVIZ program, key words appear on the map. In other systems, documents or concepts may be represented as alpha-glyphs, tadpole-like icons with attached lines or "tails" pointing in the direction of similarity.

HNC, Inc took a statistical approach. Their MatchPlus program employs neural network technology to learn database vocabulary by first discovering "similarity of usage at a word level, in a language-independent manner, without the need for external dictionaries, thesauri or semantic networks" [Caid & Carleton, 2]. The neural network generates word vectors which represent document content. Words learned in a given context point to or cluster with other related terms used in such contexts as weather, finance or government [Caid & Carleton, 3]. The next step is to represent documents in terms of "the weighted sum of the context vectors associated with words in the document" [Caid & Carleton, 5]. Documents dealing with similar topics are clustered or indexed for easier scanning. Document retrieval is accomplished by converting the user's natural language query into a context vector and searching for the nearest matching document vectors [Caid & Carleton, 7].

A statistical approach was also adopted by the National Security Agency's ACQUAINTANCE program which employs a "language-independent n-gram method of sorting and retrieving documents by language and topics. N-grams refers to "sequences of n consecutive characters" [Damashek, 39]. PARENTAGE, a visualization program, is used to explore the retrieved documents (See below).

Although not exhibited at the conference, it should be noted that GE Research and Development Center's NLDB text-based information system is perhaps the first to employ a hybrid approach to text retrieval. NLDB imitates human indexers and "automatically assigns categories to news stories for dissemination, retrieval, and browsing" [Jacobs]. Based on recall and precision criteria (retrieval of a high proportion of all relevant documents), their tests show that a combined knowledge-based and statistical approach to term categorization is superior to using either method alone.

Intuitive User Interfaces (IUI)

Regardless of the technology used to process, index and extract information from textual sources, it is the interface with which the user must interact. James A. Wise, Battelle Pacific Northwest Laboratory, stated that an IUI is only "'intuitive' to the degree that it exceeds the bounds of syntax". He emphasized that it must capture and communicate "the essence of meanings in messages through both the verbal and nonverbal domains", utilizing "analogs of the kinds of things that inform intuitions in everyday life". This interface must allow the user to instantly grasp the breadth of the information environment, easily filter out irrelevant material and focus search upon the most promising areas.

Several approaches to displaying the information domain were described above: proportional-colored boxes, concept maps and alpha-glyphs. Starfields, appearing like explosions of multicolored confetti, may be used in conjunction with a

legend at the bottom of the screen to indicate what the colors represent. In a library setting, for example, “blue” could indicate a mystery novel or film.

Viewers of the National Security Agency’s PARENTAGE program first see a screen covered with various concentrations of dots. This is the overview. To filter out some of these points, the user may click on a label in a menu beside a given dot cluster. This results in a cross-sectional view resembling a concept map in which related document nodes are linked by lines of varied thickness depending upon the strength of the relation. Document clusters may be searched by using a query template containing one or more labels and setting that label equal to a specific value, such as Profession s Engineer (See Figure 1). This results in a list of documents or a display of objects from which to make final selections [Cohen, 115].

Logicon’s BROWSER is designed for the user who says, “I don’t know what the evidence is likely to be. I’ll know it when I see it.” A document folder metaphor is used to aid researchers in grasping quickly how the system works. The user begins by looking at a list of folder names. Opening a folder will display a list of subject lines. Clicking on a subject line provides information about a specific document. The user may select a document and then click on a word highlighted in a text in order to see “a list of folders that include documents containing the term, bring up a list of documents in any folder on that list, and open any document”.

All of these approaches sounded terribly futuristic to attendees until they stepped into the exhibit hall and saw actual demonstrations of the systems discussed in the presentations.

A Paradigmatic Shift in How We View Information

For those using these new interface designs, a major shift in how we think about information is required. In the Gutenburg galaxy, information was organized linearly. Text and pictures followed in an orderly succession and was searched from beginning to end. In the electronic age, paper, pictures, movies, and sound recordings are collections of objects adrift in cyberspace which can be manipulated by individuals. Information is reduced to related data chunks which may be viewed contextually through text visualization [Hilbing, 54]. Order is initially imposed by various algorithms, saving the user great time and effort. In a way, the concept of information as inter-related chunks of data is analogous to the organization and storage of information in the human brain’s neural network hierarchies. Learning is the formation and modification of neural synapses in the process of forming more complex structures. HNC, Inc’s MatchPlus forms text associations in this manner. Visualization techniques transform associated data chunks into a cohesive visual representation that can be understood by the user.

A Paradigmatic Shift in Reference Support Services

Text visualization and associated intelligent systems will transform not only how the researcher locates information, but traditional reference support services as well. Since users will be able to search library holdings and documents on their own, only the most difficult reference work will need to be performed by staff [Hayes, 4].

Through text visualization and related technologies, researchers will be able to view vast amounts of data at a glance, focus their search and explore areas relevant to their interests. This will allow them greater time to interpret, analyze and report information than they have ever had. Reference staff will be free to assist researchers with more challenging reference problems. The intelligent technologies under development for the intelligence community today will be in our reference rooms tomorrow. While there are still problems to be solved before these systems become generally available, the day of the electronic reference room without walls is closer than we may wish to think.

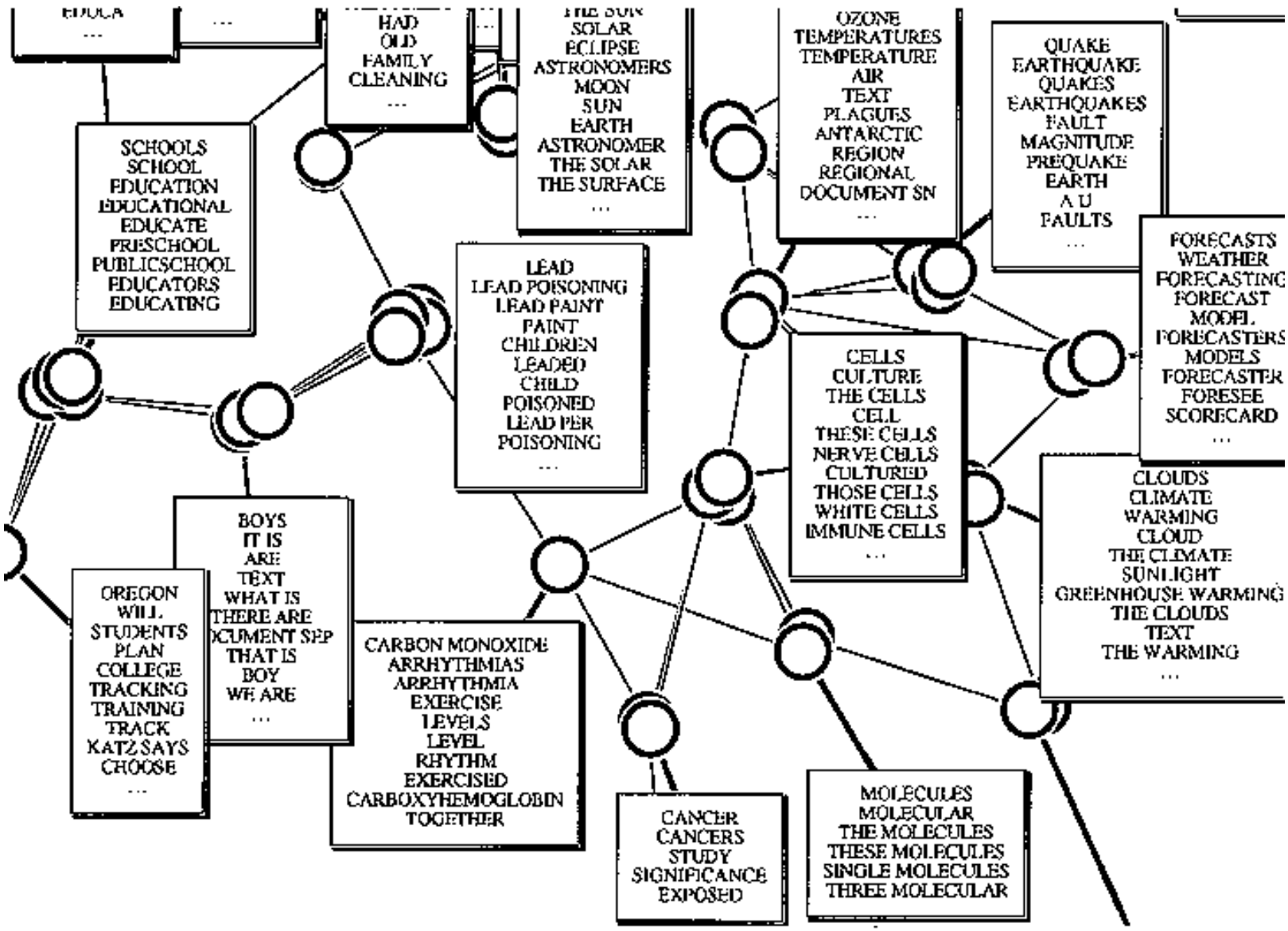

Figure 1: National Security Agency's PARENTAGE (Used with permission)

References.

Caid, William R. and Joel L. Carleton. “Context Vector-Based Text Retrieval” (HNC, Inc. nd). Working paper supplement to paper presented at Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992.

Carlotto, Mark J. “Text Visualization” (TASC, 1992). Paper presented at Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992.

Cohen, Jonathan, “Parentage”. Paper presented at Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994.

Combs, Nathan H. “Large Text Database Visualization” (TASC, 1992). Paper presented at Symposium on Advanced Information Processing & Analysis, March 24 - 26, 1992.

Croft, W. Bruce. “Knowledge-based and Statistical Approaches to Text Retrieval”, IEEE Expert (April, 1993).

Damashek, Marc. “ACQUAINTANCE”. Paper presented at Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994~-

Hayes, Phil. “Knowledge-based Systems for the Information Industry”, IEEE Expert (April, 1993).

Hilbing, Capt. John F. “Electronic Production in the Post Paradigm Shift World”. Paper presented at Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994.

Jacobs, Paul S. “Using Statistical Methods to Improve Knowledge-Based News Categorization”, IEEE Expert (April, 1993).

Meadow, Charles T. Text Information Retrieval Systems (NY: Academic Press, 1992).

Proceedings. Symposium on Advanced Information Processing and Analysis, 22-24 March, 1994. Sponsored by Advanced Information Processing and Analysis Steering Group, Intelligence Community.

Salton, Gerard. Automatic Text Processing: The Transformation Analysis, and Retrieval of Information by Computer (Reading, MA: Addison-Wesley, 1989).

Sasseen, Robert V. and William R. Caid. “Docuverse: A Context Vector-Based Approach to Graphical Representation of Information Content” (HNC, Inc. nd)

Textviz, “Visualization of Large Text Document Databases: Suggested Statement of Work for Text Visualization (TEXTVIZ) Development and Demonstration (27 February 1992). Paper accompanying presentation at AIPASG Symposium, 1994.

NOTE: Descriptions of exemplars of these systems should not be construed as endorsement of any particular system or retrieval-visualization methods by either NARA or the author. All opinions expressed are those of the author and do not reflect NARA policy or programs.

ISBD(CF) REVIEW GROUP
Meeting of April 24-26, 1995

~ Summary Report ~

John D. Byrum

The ISBD(CF) Review Group meet at the Library of Congress April 24-26, 1995 to consider a revised version of the text of the International Standard Bibliographic Description for Computer Files (1990) prepared at the chairman's request by Ann Sandberg-Fox who is serving as principal editor of the Second Edition. In attendance at this meeting were Group members Sten Hedberg (Uppsala Universitetsbibliotek); Catherine Marandas (Bibliotheque nationale de France); Ms. Sandberg-Fox (Colchester, Vermont); chairman John Byrum (Library of Congress) as well as corresponding members Laurel Jizba (Michigan State University Libraries) and Lucy Evans (British Library) as well as observer Claire Vayssade (Bibliotheque nationale de France). The meeting was made possible by a subsidy from IFLA and a grant from the Research Libraries Group (RLG).

The first day was devoted to discussion of several issues-papers which Ms. Sandberg-Fox had prepared. These covered the topics most in need of reconsideration in the light of the rapidly developing technology which has influenced the creation and dissemination of Computer Files: Interactive multimedia; the General Material Designation (GMD); Sources of information; Reproduction and multiple versions; Designation of file; and, Published versus unpublished remote texts. In addition other aspects, such as Preliminaries, Type and extent of file, Physical description and notes were thoroughly discussed, as were a number of proposals received by the chair prior and subsequent to the formation of the Review Group. On the second and third days, the members focused on a close reading of the revision prepared by Ms. Sandberg-Fox, with the result that an agreed upon text emerged from the meeting. The draft will now be updated to incorporate decisions taken at this gathering and, with permission of the Sections on Cataloguing and on Information Technology, presented for world-wide review on or about September 1, 1995. Following a six-month comment period, a final version of ISBD(CF) Second Edition will be readied for IFLA approval and publication; in addition, the text will be shared with the authors of national and international cataloguing codes, such as the Joint Steering Committee for AACR.

Following is a brief summary of the most important outcomes of the April 24- 26 meeting and which will be reflected in the revised ISBD(CF), presented in terms of the objectives that were set out to guide this project:

(1) To take into account the emergence of interactive multimedia, a new and still developing technology that combines and stores products of audio and video technologies, together with text and graphics, on optical discs.

Regarding interactive multimedia, the Review Group concluded that all such resources be incorporated into the new version of CF. This conclusion was reached because no existing ISBD covers these materials (which entered the mass market beginning in the mid-1980's), and because user-manipulated, non-linear navigation using computer-controlled technology are hallmarks which characterize interactive multimedia. (These materials are distinct from multimedia/kits that are covered by the stipulations of ISBD(NBM).) As a result, the new version of CF will add or amend provisions regarding sources of information (0.5), edition (area 2), type and extent of file (area 3), dates (area 4), physical description (area 5) and the notes (area 7) to show treatment of interactive multimedia as a subset of computer files. Examples will be added to illustrate such files.

(2) To consider the impact of developments in optical technology, as new and improved optical discs are replacing magnetic disks as primary storage devices.

The Review Group decided to improve CF to cover not only CD-ROMs (compact disc read-only memory) but also CD-I's (compact disc interactive), and other emergent forms such as photo-optical compact disc. As a result, the new version of CF will add or amend provision regarding sources of information (0.5), edition (area 2), physical description

(area 5), and notes (area 7). The term “disk” (spelled with “k”), currently used throughout area 5 to describe both optical and magnetic devices, will now apply only to magnetic devices, while “disc” (spelled with “c”) will be used in relation to optical manifestations.

(3) To provide for the availability of remote electronic files on the Internet, a global network of networks that allows users access to a vast wealth of remote electronic files, including books, journals, articles, reference sources, and even library catalogs.

Since, at the time CF was first formulated, this was a new area especially designed to treat these files, caution was exercised as to the kind and amount of detail to be given. Designations of the type of file are limited to general terms only—”Data” and “Program” and their combination “Data and program.” The Review Group decided that these terms are not adequate for the purposes of identifying the many different types of data files and software on the Internet. Indeed, the whole treatment of the Designation of file was thoroughly reworked and developed, with area 3 emerging as the one most thoroughly changed in revised CF. Consequently, the Second Edition of CF will propose several levels of specificity as appropriate. The current terms “Data” and “Program” will continue to be authorized, but Data files can alternatively be indicated as “Numeric”, “Text”, Pictorial”, “Representational” or “Sound”, while Programs can be identified as “Utility”, “Application” or “System”. Most of these categories are further delineated for more specific designation when appropriate; for example, a Bibliographic database may be so identified, as may be a Game. As before, the combination “Data and Program(s)”will continue to be used when applicable. However, alternative identification as to particular types of data and program(s) may be taken from the authorized listing and be used in conjunction with the following terms: “Interactive multimedia” or “Online service.” These latter terms also function as designations when terms from the authorized listing are not appropriate. Where, in the case of combinations, the program or the data may be incidental to the whole, the primary term only is to be given. As for the General Material Designation (GMD), the Group decided to retain “Computer File” in the absence of a better alternative.

Further addressing Internet resources, the revised CF will provide better treatment of the networking environment where an electronic file may be accessed by several methods, reside in many directories, and require more detailed information, enabling users to locate and retrieve these files. Specifically, CF will be updated to include provision for URL’s, gopher and FTP sites.

c (4) To deal with bibliographic problems arising from reproductions of computer files such that many CF titles are now available in a variety of physical formats.

Although such problems are not easily resolved, the CF Review Group did authorize changes to areas 2 and 5 to better distinguish between an “original” and other versions thereof. Reformatting changes were moved from inclusion in the definition of edition to inclusion, instead, into the definition of what would not constitute a new edition. Also, output medium and display format are newly reworked phrases to better reflect CF technology.

In addition, the Review Group agreed to significant modifications of the provisions concerning Sources of information (0.5). Area 4 (“Publication”) will be amended to require treatment of all remote CF as published materials. In addition, the glossary and examples will be updated and increased.

In the course of its meeting, as requested, the Group considered the Official Draft Proposal of the IFLA Division of Bibliographic Control Study Group on the Functional Requirements for Bibliographic Records, with Barbara Tillett, one of the three consultants to that project, present for part of the discussion. It was decided that as a medium, computer files would provide a good test of the draft, and the Group agreed to undertake an in-depth study. Specifically, 1) the use of the words “item” and “work” in the Functional Requirements document will be examined in relationship to related terminology in the ISBD(CF); 2) an experiment will be conducted to apply the suggested model using several types of computer files in several library environments; 3) the results of the experiment will be analyzed; and 4) a summary document, including any potential recommendations for the ISBD(CF) will be written. Laurel Jizba will coordinate this study for presentation by November 1, 1995.