published as: Fabrikant, S. I. and Buttenfield, B. P. (1997). Envisioning User Access to a Large Data Archive. *Proceedings*, GIS/LIS '97, Cincinatti, OH, Oct. 28-30, 1997: 686-691 (CD-ROM).

ENVISIONING USER ACCESS
TO A LARGE DATA ARCHIVE

Sara I. Fabrikant and Barbara P. Buttenfield
Department of Geography, Campus Box 260
University of Colorado, Boulder CO 80309
voice: (303) 492-3684
email: sara.fabrikant@colorado.edu and babs@colorado.edu

ABSTRACT

We are overwhelmed by the vast amounts of data accumulating daily. The extraction of information from online data sources is becoming more and more difficult. For example, if a query to a large archive returns hundreds of “hits”, the most effective presentation is probably not a list of items, but some other type of graphical display. The concept of spatialization offers a promising potential to overcome the current impediments of retrieving items from large volume archives. Spatialization involves effective combination of powerful scientific visualization techniques with spatial metaphors that represent data that are not necessarily spatial in nature. Familiar spatial concepts such as distance and direction, scale, arrangement etc. which are part of the human experience in everyday life, are applied to create lower-dimensional digital representations of complex digital data. Skupin and Buttenfield (1996;1997) have demonstrated how spatial metaphors can be constructed for abstract information spaces. However, as these authors (1996: 616) point out, there has not yet been any subject testing to determine the appropriateness of such methods for visualization. We are not certain how people comprehend spatialized views, or whether the components of distance, direction and so forth are understood by viewers. This paper presents an experimental design to explore how spatialized views are understood by users. Subject testing procedures on graphical displays are outlined, which include the collection of performance measures on information retrieval tasks. The experimental application will rely on data collected from the Alexandria Digital Library Project.

INTRODUCTION

Searching data and retrieving information from large online data archives can be a very frustrating experience. A user might encounter the following interactions with a system, during a query session of a large data repository:

> The user fills in a query form, by entering keywords in a text entry field, as well as other related information such as dates, numbers etc. to refine the search. Ideally, the system returns a small set of "hits", which are related to the search keyword and include the desired information. However, an information seeker is often overwhelmed by a huge amount of returned query results. Consequently, the user has to go through the time consuming process of sifting through large amounts of data, which might not be related to the requested search.

Yet another query might result in zero "hits", or leaves the user with the feeling of having used a wrong keyword, or having misused a query option. In both cases, the user has to refine the query, until the desired subset of items is returned by the system, thus requiring the users time and effort.

SPATIALIZATION

The concept of spatialization offers a promising potential to overcome to current impediments to efficient information processing and retrieval. Spatialization refers to the effective combination of powerful scientific visualization techniques with spatial metaphors that represent complex high-dimensional data sets, which may be non-spatial in nature. Familiar spatial concepts such as, distance, direction, scale and arrangement which are part of the human's experience in everyday life, are applied to create low-dimensional digital representations of complex digital data. As Chalmers (1993:378) puts it: "our everyday world is 2.1 dimensional, therefore physical spaces of high dimensionality are unfamiliar to most of us, and it is generally more difficult to present, perceive and remember patterns and structures within them."

The user's understanding of spatialization is based on envisioning spatial properties. Furthermore it relies on cognition of geographical space, which involves memory, spatial reasoning and communication about objects, their spatio-temporal and thematic attributes, as well as the relationships among these objects in the real world (Montello, 1996).

Golledge (1995) presents a minimal set of primitives for building spatial concepts. These include identity, location, magnitude, and time. Distance, angle and direction, connection and linkage (nearest neighbor, proximity, similarity etc.) are derived concepts from the first order primitive location. Higher order spatial concepts are combinations of derived concepts. For example, if location, magnitude, and connectivity are combined, we obtain the concept of an ordered tree, which provides a useful metaphor or data model for the concept of scale. Likewise, location may be combined with magnitude to obtain (local) density, and build up the concept of dispersion.

A very common example for the application of spatial metaphors to envision an abstract computer environment is the desktop metaphor developed by Apple as a graphical user interface for the Macintosh computer. The two dimensional view of a computer operating system as an office table, covered with folders and documents, allows one to visually collect, process and store digital data.

Using the spatial properties such as proximity, we typically regroup related files or applications, by putting them into a common folder. Consequently, hierarchies of folders can be created, to simplify navigation through "data space". Deeper into the hierarchy, more detailed information about the data is revealed, thus relating to scale dependence in the real world. Moreover, by surmounting distance with the "drag" and "drop" option, we are able to perform actions within the computing environment, such as copying or deleting files. Files which have to be deleted are carried to a specific place on the "office table", to be put into a "trash can". Typically the trash resides somewhere at the edges of the "office table", neither obstructing our working environment, nor being too close to important files.

FROM QUERIES TO BROWSING AND FILTERING

The retrieval of information from large data archives has long been an important issue in computer science (Parsaye et al, 1989). A common problem for information retrieval is related to the user interface of the query system. The user interface generally provide insufficient guidance and queries often return a huge set of undesired results (Doan et al, 1996).

The term *information retrieval* is being set aside by newer information seeking strategies, such as *data browsing, data mining, data warehousing,* or *filtering* (Shneiderman, 1996). Common to the newer information gathering terms are their exploratory nature and the integration of sophisticated direct-manipulation user interfaces, supporting what Shneiderman calls the *Visual Information Seeking Mantra.* The mantra includes three parts: "overview first, zoom and filter, then details-on-demand" (Shneiderman, 1996). In related work, Doan et al (1996) propose dynamic queries, using the direct manipulation approach, where the query process as well as the results carry a visual component. Continuous graphical feedback supports the user in query formulation and subsequent query refinement.

To design a query system based on spatial metaphors, Shneiderman would have us first define the kinds of queries a user would typically perform. We can apply these to the Alexandria Digital Library (ADL, on the Web at http://alexandria.sdc.ucsb.edu/), a distributed digital library for geographically referenced information. The schema in Figure 1 outlines how information seekers can interact with ADL's collection.

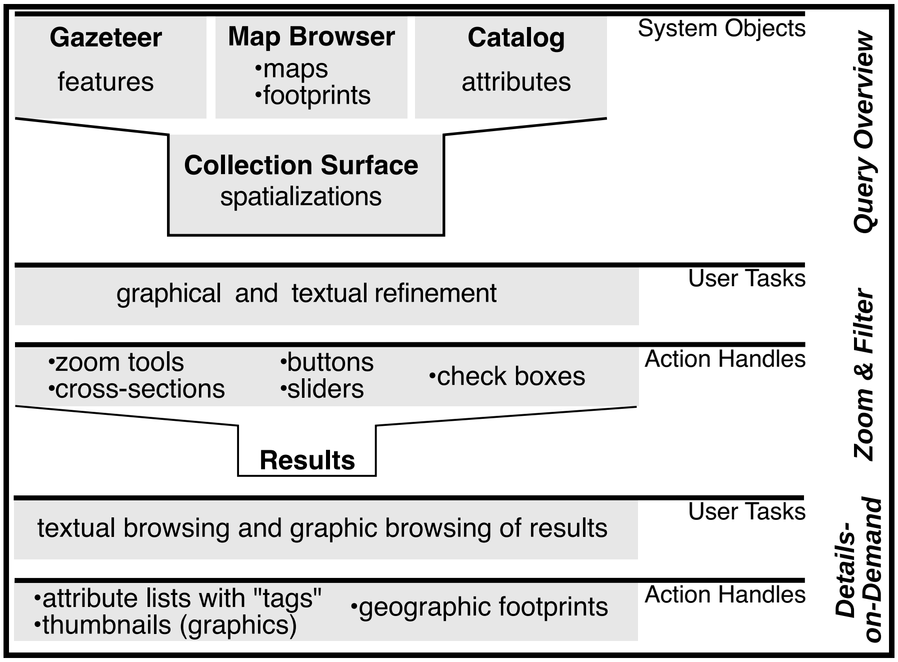


Figure 1: Visual Browsing Query Process

In the current interface, there are three ways to query the library: entering specific keywords, in the gazetteer (geographic search) or in the catalog (attribute search), or use the map browser to graphically refine the search area, by zoom and pan. In all query stages, user tasks are restricted to "known-item-searches", thus requiring specific keywords or geographic areas as query inputs.

Whereas specific fact finding will be well served by the described system, exploratory querying and open-ended browsing are not supported adequately. For example, some users might not have a well defined information need. Others might desire to gain an overview over the entire collection first, before deciding on a specific topic. Finally, information seekers might be interested in discovering relationships among the items in the database, enabling them to formulate unforeseen queries.

SPATIALIZED BROWSING IN A DIGITAL LIBRARY

"A picture is worth a thousand keywords"

Drawing upon the work of Skupin and Buttenfield (1996; 1997) who demonstrated how spatial metaphors can be constructed for abstract information spaces and Shneiderman's (1996) Visual Information Seeking Mantra a spatialized query session in ADL could be envisioned as follows:

The query process is divided into three stages: overview first, then zoom and filter, and lastly, details-on-demand. The graphical user interface (GUI) for the overview stage is a direct manipulation interface with linked windows (Figure 2). Dynamic queries are carried out by buttons, sliders and check boxes, which trigger an immediate graphical response by the system. Items selected in the lists will be highlighted in the spatialized views and visa-versa.

Three spatial metaphors underlie the design of this graphical user interface, including distance (similarity), scale (level of detail), and arrangement (dispersion and concentration).

Distance

In Figure 2, the large window displays a landscape of catalog items that were "hit" by a query. Items that are close together are characterized by similar keyword sets. In the abstract data space we may interpret distance as similarity in a metaphoric sense. Catalog items which are more related to each other will be placed closer together than items which are less related. The distance metaphor is based on Salton's (1989) vector space model (keyword occurrences in a document), and multidimensional scaling (MDS) is utilized as the projection method (Skupin and Buttenfield, 1997).

Scale

The interface has several components designed to make the level of detail evident to the user. Keywords can be selected in a Hierarchy Tree Window, which will update other display windows accordingly. In the Figure, selecting the keywords 'aerial photograph' and 'cartographic material' highlight the same keywords in the Hierarchy Tree Window and the Keyword List Window. Keywords can be tagged and the selected items can be promoted to the top of the list. As check boxes are tagged, the Landscape Window is updated to display the keyword labels on the landscape.

In the lower left corner of the Figure, a window reacts dynamically to keyword selection by displaying bars showing the relative percentage of "hits" that would be associated with each keyword in the collection. The Cross Section Window represents a frequency of "hits" that could be expected by refining the query as defined by the transect line drawn in white across the Landscape Window. These windows operate together to help the user predict the probable success rate for a given query as it is formulated. A tool palette over the Landscape Window allows 'zooming in' on the data space to see the landscape (and information about the collection) in more detail. Zooming tools also modify the Keyword List and Hierarchy Tree Windows.

Arrangement

The Landscape Window in Figure 2 is a "collection surface" which offers a visual overview of queried items. The z-values in the landscape represent the accumulated number of hits per "region" in the collection. A high peak indicates a high concentration of items available for that particular query. Patterns and shapes in the landscape reveal the organization of items with respect to each other. For example a steep cone indicates a high density of similar documents correlated with a high number of hits. A low plateau on the other hand describes a lower density of items available.

The Map, Catalog, and Gazetteer check boxes re-arrange items in the Landscape Window to identify catalog keywords, map browser footprints, and gazetteer features, respectively. Whereas the first representation, based on geographic regions, is well known to the geographic information community, the abstract keyword landscapes are not as familiar. Chalmers (1993), Atkins (1995), and Skupin and Buttenfield (1996; 1997) have shown how effective spatial metaphors can be utilized to construct abstract information landscapes.

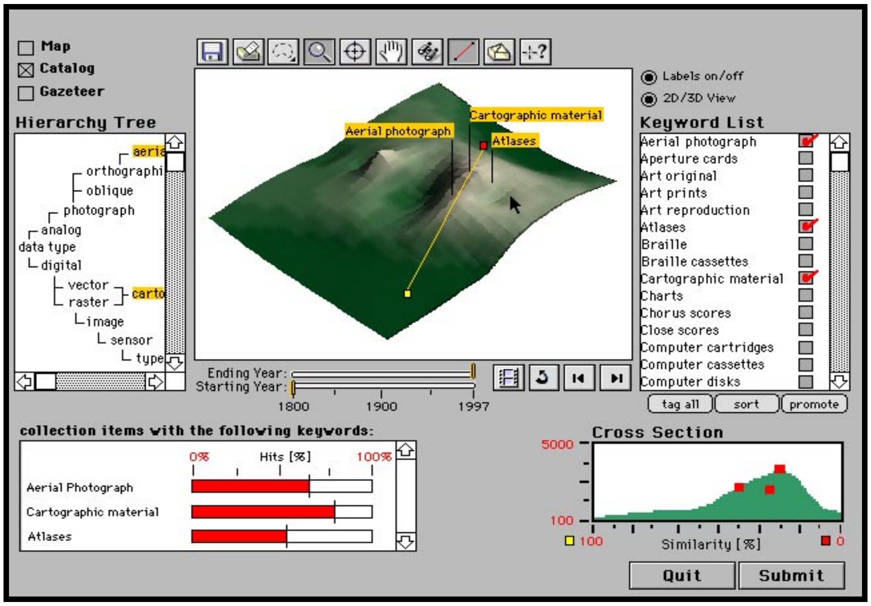


Figure 2: User Interface for the Overview Query Stage

EVALUATION OF SPATIAL METAPHORS

Skupin and Buttenfield (1996: 616) point out, that there has not been any subject testing to determine the appropriateness of the described methods for visualization. We are not certain how people comprehend spatialized views, or whether the components of distance, direction and so forth are understood by viewers. Spatializations rely on the use of spatial metaphors to represent data that are not necessarily spatial. Metaphors constitute a fundamental part of human cognition (Lakoff, 1987). Lakoff (1987) defines the Spatialization of Form hypothesis, which requires a metaphorical mapping from physical space into a “conceptual” space. Consequently, image schemata which structure space are mapped into the corresponding abstract configurations, which structure concepts (i.e. similarity) (Lakoff, 1987: 283). To inquire how well the metaphorical mapping is assimilated by a user leads us to the main research question: What kinds of skills are needed to understand the spatializations?

Metaphor	Question
Distance	How well do people understand the concept of similarity?
Scale	Can people discern hierarchical order?
Arrangement	Can people detect regions in the display?

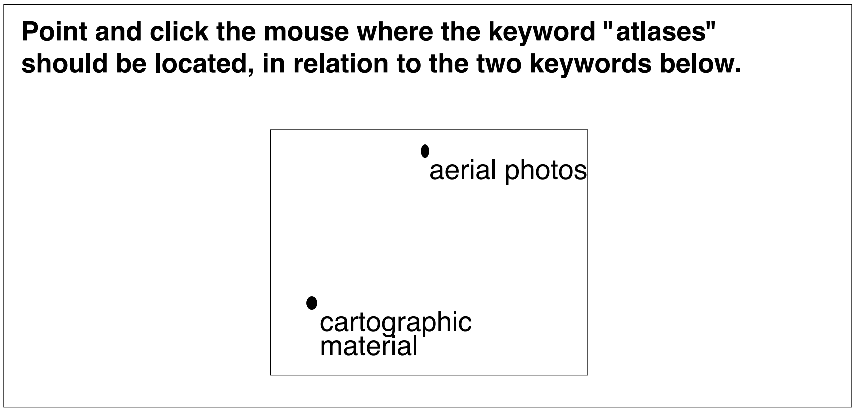
Table 1: Research questions

Distance

A way to test this metaphor is by using the technique of comparative distance judgment tasks. Consequently, the complete method of triads is used to obtain comparative distances between stimuli (Torgerson, 1958). The judgment tasks are presented in triads, in the form: "the keyword *atlases* is more similar to the keyword *cartographic material* than to the keyword *aerial photographs"*. To extract all relationships between the three stimuli, three questions have to be asked, giving a triadic combination. Thus, with *n* stimuli there are:

$$\frac{n(n-1)(n-2)}{6} \text{ triads and } \frac{n(n-1)(n-2)}{2} \text{ judgments} \tag{1}$$

for each subject. From these judgments we obtain the proportion of times any stimulus x is judged more similar to stimulus y than to z. For example, test subjects are presented with triads in the form:



Point and click the mouse where the keyword "atlases" should be located, in relation to the two keywords below.

Figure 3: Subject Test for Distance Between Keywords

The obtained proximity matrices are subjected to a multidimensional scaling algorithm. The resulting test data space is then overlaid onto the keyword vector space model to produce the spatialized view. The comparisons of the two spatializations could provide further insights.

Scale

When examining the scale metaphor, we want to inquire how well users comprehend hierarchical order in the data archive. Hierarchy is composed of the spatial primitives identity, location, magnitude and connection. In Lakoff's (1987) terms, hierarchy is an example of the *part-whole schema*. A sample test for this metaphor is to present test subjects with a set of stimuli, such as keywords from the database, and ask them to group the stimuli according to their rank in the hierarchy. The obtained hierarchies are then compared with the existing hierarchical order in ADL. One could also utilize cluster analysis to validate user choice. Utilizing the spatialized displays, test subjects have to use the zoom tool several times, and indicate which of the presented hierarchical keyword lists, match the keywords displayed during the zoom.

Arrangement

Spatial distribution includes areal pattern types, such as dispersion and density. Density is unit dependent. The unit in a data archive is a document, represented by a keyword. In the three dimensional case, magnitude is added to the spatial distribution, giving the number of "hits". Consequently, peaks of concentration and valleys of dispersion indicate spatial distribution on the collection surface. The question arises, if users can discern regions in the spatializations and to which extent the concept of concentration and dispersion is understood.

Tests for the arrangement metaphor include a spatialized display. Subjects are asked to "lasso" an area which corresponds best to a given keyword. Subjects are asked to place a mark in the display, where minimum and maximum concentration occur. Both correctness of the answer and the speed of response are measured.

Higher order derived concepts such as gradient or slope can be tested as well. As the number of hits vary over the collection surface, the slope increases either sharply or in a more gentle fashion. A steep slope indicates a short distance between two points, as well as an abrupt change in magnitude. In other words, documents are represented with very similar content, but with distinctly different numbers of "hits". The combination of similarity versus frequency is tested with a profile display. Test subjects have to select the appropriate statement, which best represents a section of the profile.

SUMMARY AND OUTLOOK

The use of spatial primitives to query a large data repository has been outlined. A spatialized graphical user interface has been presented, which allows the exploration of the holdings of the Alexandria Digital Library. Although the concept of spatialization is not entirely new to the research community, and several authors have demonstrated how spatial metaphors can be constructed for abstract data spaces, it's appropriateness for visualization has not been tested yet. We have outlined what kinds of questions have to be asked, to reveal if spatial primitives such as distance, direction and scale are understood by viewers of spatialized displays. There is an imminent need for empirical evaluation and validation of emerging procedures and techniques in the visualization domain. The geographic information science community, with it's wealth of experience in spatial information processing, is predestined to add valuable insights to the spatialization domain. The results of this research should fuel the enormous potential spatialization has to offer, to overcome the bottleneck of information processing.

ACKNOWLEDGEMENTS

This paper forms a portion of the Alexandria Digital Library Project, jointly sponsored by NSF, NASA, and ARPA. Funding by the National Science Foundation (NSF IRI-94-11330) is greatly appreciated. Matching funds from the University of Colorado are also acknowledged.

REFERENCES

Atkins, P. W. (1995). *The Periodic Kingdom.* Basic Books, New York.

Chalmers, M. (1993). Using a Landscape Metaphor to Represent a Corpus of Documents. In: Frank, A. U., Campari, I. (Eds.). *Spatial Information Theory. A Theoretical Basis for GIS. Lecture Notes in Computer Science,* No. 716, Springer, Berlin: 377-390.

Doan K., Plaisant C., and Shneiderman B. (1996). Query Previews in Networked Information Systems. *Proceedings of the Third Forum on Research and Technology Advances in Digital Libraries, ADL '96*, Washington, DC, May 13-15, 1996, IEEE CS Press: 120-129.

Golledge, R. (1995). Primitives of Spatial Knowledge. In: Nyerges, T. L., Mark, D. M., Laurini and R., Egenhofer, M. J. (Eds.) *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems,* Kluwer Academic Publishers, Dordrecht: 29-44.

Lakoff, G. (1987). *Women, fire and dangerous things. What categories reveal about the mind,* University of Chicago Press, Chicago.

Montello, D. R. (1996). *Cognition of Geographic Information.* Research Priorities for Geographic Information Science. Univeristy Consortium for Geographic Information Science, Paper #4, http://www.ncgia.ucsb.edu/other/ucgis/research_priorities/paper4.html (July, 1997).

Parsaye, K., Chignell, M., Khoshafian, S., and Wong, H. (1989). *Intelligent Databases. Object-Oriented, Deductive Hypermedia Technologies.* Wiley, New York.

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

Skupin, A. and Buttenfield, B. P. (1997). Spatial Metaphors for Display of Information Spaces. *Proceedings, AUTO-CARTO 13,* Seattle, Washington, Apr. 7-10, 1997: 116-125.

Skupin, A. and Buttenfield, B.P. (1996). Spatial Metaphors for Visualizing Very Large Data Archives. *Proceedings, GIS/LIS '96,* Denver, Colorado, Nov. 19-21, 1996: 607-617.

Shneiderman, B. (1996). The Eyes Have It. A Task by Data Type Taxonomy for Information Visualizations. *IEEE Symposium on Visual Languages 1996, Proceedings,* Boulder, CO, Sep. 3-6, 1996: 336-343.

Torgerson, W. S. (1958). *Theory and Methods of Scaling.* Wiley, New York.