# Evaluating User Requirements for a Digital Library Testbed

Barbara P. Buttenfield
NCGIA
Department of Geography, 105 Wilkeson
SUNY-Buffalo, Buffalo, NY 14261

GEOBABS@UBVMS.CC.BUFFALO.EDU

## ABSTRACT

The development of widespread capabilities for electronic archival and dissemination of data can be coupled with advances in information systems technology to deliver large volumes of information very fast. Paradoxically, as greater volumes of information become available on electronic information networks, they become increasingly difficult to access. Nowhere is this situation more pressing than in the case of spatial information, which has been traditionally treated as a 'separate' problem by archivists, due to complexities of spatial ordering and indexing. A research project recently funded by NSF will address these problems and implement a working digital library testbed over the next four years. This paper will focus upon one aspect of the testbed, namely evaluating user requirements to inform interface design. The paper presents the evaluation plan, using hypermedia tools to collect real-time interactive logs of user activities on the testbed under design. Conventionally, interactive logging is analyzed by deterministic measures of performance such as counting keystrokes. In this project, the interactive log data will be analyzed using protocol analysis, which has been shown to provide a rich source of information to formalize understanding about semi-structured and intuitive knowledge.

## INTRODUCTION

The development of widespread capabilities for electronic archival and dissemination of data can be coupled with advances in information systems technology to deliver large volumes of information very fast. As greater volumes of information become increasingly available on electronic information networks, they become increasingly difficult to access. This paradox calls for research to implement intelligent software that provides and preserves access to electronic information, creating a digital library for bibliographic and analytical use (Lunin and Fox, 1994). A research project recently funded by NSF (the Alexandria Project) will address these problems and implement a working digital library testbed over the next four years. The project requires assessment of user needs, basic research to address technical impediments, software development, and a rigorous program of evaluation and quality control.

This paper will focus upon one important aspect of the testbed, namely evaluating user requirements to inform interface design. Challenges associated with interface design for distributed data have been reviewed in Kahle et al (1994). This paper argues for a return to empirical evaluation of GIS interfaces using alternative paradigms to the psychophysical designs once popular in cartographic research.. A brief chronology of empirical evaluation research in cartography is presented. Semi-structured interviews are presented as an alternative paradigm for eliciting cartographic knowledge. This is followed by a description of the Alexandria Project, emphasizing user requirements and evaluation.

The paper will present the user evaluation plan, which involves real-time interactive logging of user activities using hypermedia tools to simulate the look-and-feel of the testbed under design. Conventionally, interactive logging is analyzed by deterministic measures of performance such as counting keystrokes, recording time units between system dialog and user response, etc. These types of analyses are useful but limited in their neglect of user cognition. Interactive user logs collected for this project will be

analyzed using Protocol Analysis, which has been shown to provide a rich source of information to formalize understanding about semi-structured and intuitive knowledge (Ericsson and Simon 1993). Its application to interactive logs has not been reported in the literature.

## EMPIRICAL EVALUATION IN CARTOGRAPHIC RESEARCH

Empirical research has largely disappeared from the cartographic literature following dissatisfaction with the psychophysical methods that figured prominently for several decades (roughly, 1965-1985). To be clear, the dissatisfaction lay not with the analytical methods, which tend to be highly structured, highly deterministic, and highly confirmatory. The dissatisfaction lay rather in the limited gains in understanding intermediate stages in the process of cartographic communication, meaning those stages of reasoning that occur intermediate between identifying visual elements (seeing map items) and reaching an interpretation (inference of pattern). Similar concerns were expressed at this time in other disciplines.

> After a long period of time during which stimulus-response relations were at the focus of attention, research in psychology is now seeking to understand in detail the mechanisms and internal structure of cognitive processes that produce these relations... This concern for the course of the cognitive processes has revived interest in finding ways to increase the temporal density of observations so as to reveal intermediate stages of the processes. Since data on intermediate processing are costly to gather and analyze, it is important to consider carefully how such data can be interpreted validly, and what contribution they can make to our understanding of the phenomena under study. (Ericsson and Simon 1983, p.1)

In cartography, the stimulus-response paradigm brought forth one insight consistently highlighting the sensitivity of experimental results. Estimates of symbol size were shown to vary with symbol clustering and with inclusion of basemap enumeration units (Gilmartin, 1981; Slocum, 1983). Estimates of grayscale were shown to be dependent upon screen texture (Castner and Robinson, 1969; Kimerling, 1975; Leonard and Buttenfield, 1989) as well as the visual ground on which figures appeared (Kimerling, 1985). Color identification has been shown to vary with respect to adjacent colors (Brewer, 1994) and even with viewer expectations that are unrelated to map attributes (Patton and Crawford, 1979). Several articles (Gilmartin, 1981b; Shortridge and Welch, 1980) summarize some of these sensitivities, and articulate the concern of the discipline during this time.

It is unfortunate that after this period many cartographers turned away from empirical evaluation research (but there are exceptions, for example Lloyd, 1994). For a time the published results of map evaluation were trivialized in passing remarks that cartographic researchers had become obsessed with the question "How big IS that graduated circle?". Although a body of research on interface design (for example, Laurel, 1990), on spatial conceptualization and spatial reasoning (for example, Egenhofer, 1992; Golledge, 1991; Lloyd, 1994) has been published, empirical user evaluation is pursued largely outside the GIS discipline (see for example an excellent brief review in Nyerges, 1993).

Concurrent developments in GIS software and user interfaces have not been supported with empirical evaluation. Many user requirements studies precede GIS system development (Calkins and Obermeyer 1991), however there is little or no formal evaluation once system components are implemented. System refinements are often directed towards improving performance and efficiency rather than system use. One reason for this may relate to the complexity of GIS system use. Nyerges (1993, p.48) comments "GIS applications tend to be quite involved ... in rather rich problem settings. Hence realistic task models will most likely be rather complex. Real world analyses of GIS use tasks are needed, in addition to the selective focus of controlled experiments in the laboratory."

Clearly, there is more to knowing how people come to understand what they experience than can be provided by strict adherence to models of stimulus-and-response. In the absence of adopting novel paradigms for evaluative research, user evaluation studies in cartography have languished.

## AN ALTERNATIVE PARADIGM FOR USER EVALUATION

The foundation of evaluation research is observation. Observational data analysis in the stimulus-response paradigm is documented by magnitude estimation, determination of equivalence or difference, or other measures. Most often these are metric, and their goal is to uncover perceptual patterns such as visual clustering, contrast, brightness, size, and so on. Other types of observational data analysis capture higher order cognitive patterns, such as selecting one path of action from several options, recounting steps taken to complete a process, and verbalizing a set of criteria used to solve a problem. Associated data measures are less deterministic, less structured, and must be elicited using some degree of introspection. Ericsson and Simon (1983, p.48-62) recount an extensive history establishing their position that semi-structured introspective reports including think-aloud and immediate retrospective reports can in fact reflect intermediate cognitive processes, although they caution (p. 56, 62) on the difficulty of evaluating un-structured introspection (eg. free association).

Several observational methods for evaluating semi-structured information have been presented in the literature (Sanderson and Fisher, in press). Content analysis (Krippendorf, 1980) tends to focus (as its name implies) on the substantive meaning or content of user responses. Sense-making analysis (Dervin, 1983) is used to facilitate user reconstruction of a procedural task from memory. Protocol analysis (Waterman and Newell, 1971; Ericsson and Simon, 1984, 1993) has been chosen for this project as it is often used to study an ongoing process and is of relevance to evaluating computer interface usage. It is of special utility for studying the intermediate stages of decision-making, and for eliciting knowledge about a decision while subjects are involved in a decision-making process.

"Protocol analysis ... is a common method by which the knowledge engineer acquires detailed knowledge from the expert. A 'protocol' is a record or documentation of the expert's step-by-step information processing and decision-making behavior." (Turban, 1990, p. 462) Protocol analysis was developed as a systematic methodology designed to treat semi-structured behavior and dialogs as quantitative data.

Sanderson et al. (1989) report uses for their semi-automated protocol analysis software (SHAPA) to analyze problem-solving styles of doctors and nurses, crew communication during military surveillance missions and navigational tasks, program debugging in robotics, and elicitation of domain knowledge in industrial management. A more recent version of this software (MacSHAPA) has just been released (Sanderson, 1994) which extends previous analyses and incorporates real-time video control.

Examples applying Protocol Analysis to spatial behavior and human-computer interaction can also be identified. Lewis (1982) reported on the use of 'think aloud' Protocol Analysis in support of computer interface design. Mack et al. (1983) report on a study of word-processor interface design using Protocol Analysis. Golledge et al. (1983) connected results of Protocol Analysis to a theory of children's spatial knowledge acquisition. Lundberg (1984, 1989) used Protocol Analysis to analyze various types of marketing and consumer behavior. Sumic (1991) used Protocol Analysis in dissertation research to elicit expert knowledge from a utility company's electrical engineers, to support the development of a knowledge-based system linked to ARC/INFO. The company (Puget Sound Power and Light ) was impressed enough with his work that he was hired full-time to continue his research and maintain the company's knowledge base. He reported on this work at an ARC/INFO workshop on expert systems attended by the author (Sumic, personal communication, 1991). Finally, Gould (1993) used SHAPA to analyze geographic problem-solving using maps of Puerto Rico.

**Project Overview**

As stated above, the adoption of distributed data archival and retrieval can deliver large volumes of information very fast to any user on the network. Under these conditions, procedures for organizing, browsing, and retrieving such information become increasingly complex. Nowhere is this situation more pressing than in the case of spatial information, which has been traditionally treated as a 'separate' problem by archivists, due to complexities of data volume, spatial ordering, indexing, and spatial and temporal autocorrelation. For this reason, many libraries separate their text and literary archives from map archives, effectively prohibiting the cross-referencing of literary with graphical holdings. NSF, ARPA and NASA recently issued a collaborative solicitation to provide $24 million in research funds for six projects to develop "digital libraries", software testbeds demonstrating intelligent browsing and retrieval methods for digital data of any kind. One of the six awards was made to a research team including all three sites of the NCGIA.

The Alexandria Project will create, implement, and evaluate a distributed, high performance system for examination, retrieval, and analysis of spatial information from digital collections. A major goal is to remove distinctions between text and spatial data archives (or at least make those distinctions transparent to users). Alexandria will continue for the coming four years, building its testbed in two stages. The first stage will be a prototype based on commercial off-the-shelf software, to be completed within the first year of the project. In the second stage, development of the testbed will proceed in parallel with user evaluation studies to inform system engineers and designers.

**Details and Plans for User Evaluation**

As with any software engineering problem, understanding user requirements is of primary importance to build an effective user interface (Fox et al, 1994; Laurel, 1990). Evaluation of the Alexandria testbed will include several classes of users, including those familiar with the testbed contents (geographers, earth and space scientists, and professionals in public and private sector who work with spatial data) and those familiar with library cataloging and indexing systems (data archivists, research librarians, map librarians, and so on). Either class of users can be characterized as people whose knowledge of either the geographic domain or the archival domain is deep, but whose interest in learning system architecture or command structure may be minimal. For example, a query by the geographic user class might focus upon browsing satellite imagery to learn more about deforestation within a fixed distance of a river channel over several rainy seasons. A query in the archival user class might focus upon browsing through recent map acquisitions to determine if new editions exist for a given map sheet. Evaluation testing will be performed using both classes of potential users. A third class of users will be considered to include system designers, characterized as having a deep interest in geographic content, archival, AND system design. Evaluation testing will be applied to this class as a control group.

One function of the evaluation process is to determine whether characterization of user requirements is appropriate. Initial requirements are straightforward. For successful access and query of maps and satellite data, users will require individual image/map sheet control for custodial functions (acquisition of data), cataloging and index control (collection maintenance), and bibliographic control (for research and map making). A workable spatial data browsing system will require functions supporting georeferencing and fusion of multiple data types, at multiple spatial, spectral and temporal resolutions. Some requirements are important to one but not both classes of testbed users.

It is probable that access to the testbed and to the data will modify user requirements. A second function of the evaluation process is to determine and track such changes. Interface design and evaluation must be fluid and dynamic. Repetitive testing protocols will be developed to address these issues. That is, interface evaluation will begin with completion

of the first stage Alexandria prototype. Evaluation will include questions targeting possible changes in user requirements, necessitating re-testing of some but not all test subjects. Results of the first round of evaluation will be returned to system designers to guide refinement and revision of the interface. As each version of the revised interface is completed, its components will undergo empirical evaluation, with results of the analysis informing subsequent revision. Three or four cycles of evaluation are planned through the course of the project.

A third function of the evaluation process relates to metadata and data quality. Users will need to know the reliability of information returned on queries, thus the evaluation must capture user confidence as well as user satisfaction. The bottom line of this part of the evaluation process must answer two questions. The questions are first, "Does the user get the information as requested?" and second, "Does the user avoid information which is not needed?" Answers to these may not be straightforward in all situations, nor for all three user classes, which will challenge the evaluation and analysis.

Data collection for the evaluation will be accomplished by videotaping user behaviors, by interactive logging of keystrokes and response times, and by semi-structured interview, in an electronic version of think-aloud reporting. Videotaping (direct observation) is intended to capture nonverbal responses and to provide insights about user learning styles, frustration and fatigue, for example. Interactive logging will provide quantitative data measuring response times, keystrokes and mouse activity, and indicate aspects of user and system performance. Semi-structured interviewing will provide information on user requirements and on user confidence and satisfaction levels. Testing will proceed early on at the testbed development site, at UC-Santa Barbara, and proceed from there to otehr sites designated as Alexandria partners. These partners include federal and public libraries, including Library of Congress, USGS Headquarters Libraries in Reston Virginia, and the St. Louis Public Library, and various academic map libraries around the nation. Some testing will be scheduled at various GIS conferences, human-computer interface conferences, and library science conferences planned for the coming four years.

The actual mechanism for evaluation will be an operational but simulated interface, in early versions of the system. Screens with look-and-feel capabilities identical with the testbed will simulate the user interface screens and functions. Data subsets will be embedded to experiment with query and response functions for limited portions of the spatial archive. Running 'underneath' the simulated interface will be a set of interactive logging mechanisms recording keystrokes and mouse placement, documenting response times, etc. The logging mechanism will include a dialog function to converse with the user in semi-structured question-answer mode. The dialog function will be triggered throughout the evaluation session, either by the system or by the user. System triggers will initiate dialogs during specific tasks (during file retrieval for example, or on completion of a query formation). User inactivity over a threshold timeframe will also trigger a system dialog, to ask the user about confusion, or task success, for example. Users will be able to initiate dialog as well, and encouraged to keep a journal of their activities and impressions. These dialogs will be saved in a relational database for subsequent processing by Protocol Analysis.

"Protocol analysis is notoriously difficult and time-consuming to perform." (Sanderson et al. 1989, p. 1275). Application of protocol analysis involves the following steps: choice of an encoding vocabulary, definition of encoding rules, encoding of data using Protocol Analysis software, and inter-coder reliability checks. The vocabulary provides components on which user behaviors and dialogs may be categorized, and the encoding rules impose those categorizations on the data. The analysis is conceptually similar but not identical to principal components analysis, in the sense of looking for underlying patterns, rather than in the sense of a data reduction technique. Key patterns will appear consistently in one or more user classes, and inform system designers about how interface components are used by specific user classes, under what circumstances, and conditions and/or user functions that are confusing or problematic. Key patterns identified in early iterations of the user evaluation will be applied and revised for later evaluation experiments.

The software to be utilized for analyzing the Alexandria data is MacSHAPA (Sanderson, 1994) described earlier in the paper. This software can link the dialog protocols with videotape excerpts, which will tie a record of nonverbal with verbal reports of interface use. The final step in Protocol Analysis involves inter-coder reliability checks to determine if the components and categories have been applied consistently to all test subjects. Inter-coder reliability checks insure objectivity and repeatability of analytical results. "Careful protocol analysis is time-consuming, and extensive analyses require automatization. A considerable increase in objectivity may occur, since the analysis will be accomplished with determinate rules..." (Waterman and Newell, 1971, 285).

## SUMMARY

Efforts to evaluate user activity and user interface design have been largely absent from cartographic and GIS research, and it is argued this is due in part to dissatisfaction with results of research based on the stimulus-response paradigm once popular in cartography. Alternative paradigms based on observational data analysis can provide a data source for studying higher level cognitive processes involved with learning a GIS interface, including user confidence and satisfaction. One example of this type of paradigm is Protocol Analysis, which will be adopted to evaluate the user interface for the Alexandria Project, a software testbed for intelligent browsing of distributed digital spatial databases. The testbed will be under development for the coming four years, and interface evaluation studies will run concurrent with system development. The evaluation plans have been outliend in the paper, and include videotaping (direct observation) linked with interactive logging techniques underlying a simulated system interface. The logging techniques will include not only the customary deterministic measures (counting keystrokes, etc.) but also incorporate the semi-structured dialog and interview methods practiced in Protocol Analysis. Three classes of users will be evaluated, including users of spatial data (geograpy and earth science professionals), archivists of spatial data (library and information management professionals), and system designers and engineers.
It is felt that such evaluation can only improve the flexibility of system interface design, and additionally assist researchers in formalizing some types of cartographic knowledge.

## ACKNOWLEDGEMENTS

This research forms a portion of the Alexandria Project, funded by the National Science Foundation (NSF contract IRI 94-11330), and a portion of Research Initiative 8, "Formalizing Cartographic Knowledge", of the National Center for Geographic Information and Analysis (NSF contract SBR 88-10917). Funding by NSF is gratefully acknolwedged.

## REFERENCES

Brewer, C.A. 1994 Guidelines for Use of the Perceptual Dimensions of Color for Mapping and Visualization. **Proceedings,** IS&T/SPIE Symposium on Electronic Imaging Science and Technology, Color Hard Copy and Graphic Arts III Technical Conference, 8 February 1994, San Jose, California (no page nos. on manuscript).

Calkins, H.A. and Obermeyer, N.A. 1991 Taxonomy for surveying the Use and Value of Geogrpahical Information. **international Journal for Geographic Information Systems**, vol. 5(3): 341-351.

Castner, H.W. and Robinson, A.H. 1969 Dot Area Symbols in Cartography: The Influence of Pattern on Their Perception. **Technical Monograph No. CA-4.** Washington D.C: American Congress on Surveying and Mapping.

Dervin, B. 1983 An Overview of Sense-Making Research: Concepts, Methods, and Results to Date. Presented International Communication Association , Dallas Texas, May 1983 (no page numbers on manuscript).

Dervin, B. and Nilan, M.S. 1986 Information Needs and Uses. In M.E. Williams (Ed.) **Annual Review of Information Science and Technology**, vol. 21: 3-33.

Ericsson K.A. and Simon, H.A. 1993 **Protocol Analysis: Verbal Reports as Data.** Cambridge Mass: MIT Press (2nd Edition).

Ericsson K.A. and Simon, H.A. 1984 **Protocol Analysis: Verbal Reports as Data.** Cambridge Mass: MIT Press (1st Edition).

Fox, E.A., Hix, D., Nowell, L.T., Brueni, J., Wake, W.C. Heath, L.S., and Rao, D. 1994 Users, User Interfaces, and Objects: Envision, a Digital Library. **Journal of the American Society for Information Science**, vol.44(8): 480-491.

Gilmartin, P.P. 1981a Influences of Map Context on Circle Perception. **Annals of the Association of American Geographers**, vol 71: 253-258.

Gilmartin, P. P. 1981b The Interface of Cognitive and Psychophysical Research in Cartography. **Cartographica**, vol.18(3): 9-20.

Golledge, R.G. 1991 The Conceptuall and Empirical Basis of a General Theory of Spatial Knowledge. In Fischer, M.M. Nijkamp, P. and Papageorgiou, Y.Y. (eds.) **Spatial Choices and Processes.** Amsterdam: North Holland: 147-168.

Golledge, R. G., Smith, T. R., Pellegrino, J. W., Doherty, S., Marshall, S. P. and Lundberg, G., 1983 The acquisition of spatial knowledge: Empirical results and computer models of path finding processes. Presented at XIX Inter-American Congress of Psychology, Quito, Ecuador, July 1983 (no page numbers on manuscript).

Gould, M. D., 1993 **Map Use, Spatial Decisions, and Spatial Language in English and Spanish.** Unpublished Ph.D.dissertation, Department of Geography, State University of New York at Buffalo.

James, J. M. and Sanderson, P. M., 1991 Heuristic and Statistical Support for Protocol Analysis with SHAPA Version 2.01. **Behavior Research Methods, Instruments and Computers**, vol. 23(4): 449-460.

Kahle, B., Morris, H., Goldman, J., Erickson, T., and Curran, J. 1994 Interfaces for Distributed Systems of Information Servers. **Journal of the American Society for Information Science**, vol.44(8): 453-467.

Kimerling, A.J. 1985 The Comparison of Equal Value Gray Scales. **The American Cartographer,** vol. 12(2): 132-142.

Kimerling, A.J. 1975 A Cartographic Study of Equal Value Gray Scales for Use with Screened Gray Areas. **The American Cartographer,** vol. 2(2): 119-127.

Krippendorf, K. 1980 **Content Analysis: An Introduction to its Methodology**. London: SAGE Publications, vol. 5 (The CommText Series).

Laurel, B. (ed.) 1990 **The Art of Human-Computer Interface Design**. Reading, Massachusetts: Addison-Wesley.

Lewis, C. H., 1982 Using the "Thinking Aloud" Method in Cognitive Interface Design. **IBM Research Report** RC-9265. Yorktown Heights, NY: T.J. Watson Research Center.

Lloyd, R. 1994 Learning Spatial Prototypes. **Annals of the Association of American Geographers**, vol. 84(3): 418-439.

Lundberg, G., 1984 Protocol analysis and spatial behavior. **Geografiska Annaler,** vol. 66B(2): 91-97.

Lundberg, C. G., 1989 Knowledge acquisition and expertise evaluation. **The Professional Geographer,** vol. 41(3): 272-283.

Lunin, L. F. and Fox, E.A. 1994 Perspectives on Digital Libraries. **Journal of the American Society for Information Science**, vol.44(8): 441-445.

Mack, R. L, Lewis, C., and Carroll, J., 1983 Learning to Use Word Processors: Problems and Prospects, **ACM Transactions on Office Information Systems,** vol.1(3): 254-271.

Sanderson, P.M. 1994 **MacSHAPA**. Version 1.0.2. Champaign-Urbana, Illinois: Department of Mechanical Engineering, Univeristy of Illinois. Distributed by CSERIAC, Wright-Patterson Air Force Base, Ohio.

Sanderson, P.M. and Fisher, C. (in press) Exploratory Sequential Data Analysis: Foundations. **Human-Computer Interaction**, vol.9(3).

Sanderson, P. M., James, J. M., and Seidler, K. S., 1989. SHAPA: an interactive software environment for protocol analysis. **Ergonomics** vol. 32(11): 1271-1302.

Shortridge, B.G. and Welch, R.G. 1980 Are We Asking the Right Questions? Comments on Cartographic Psychophysical Studies. **The American Cartographer**, vol.7: 19-23.

Slocum, T.A. 1983 Predicting visual Clusters on Graduated Circle Maps. **The American Cartographer**, vol. 10(1): 59-72.

Williams, R.L. 1958 Map Symbols: Equal Appearing Intervals for Printed Screens. **Annals,** Association of American Geographers 48:132-139.

Turban, E., 1990 **Decision Support and Expert Systems: Management Support Systems**. New York: Macmillan.

Waterman, D. A. and Newell, A., 1971 Protocol Analysis as a Task for Artificial Intelligence. **Artificial Intelligence** vol.2: 285-318.