Evaluating the Usability of the Scale Metaphor for Querying Semantic Spaces

Sara Irina Fabrikant

Department of Geography
University of California Santa Barbara
Santa Barbara, CA 93106-4060 USA, sara@geog.ucsb.edu

Abstract. Information visualizations have become popular tools for extracting knowledge from large bodies of information. Very little is known on the usability of such 'visual knowledge tools' for information access. The goal of this paper is to show the usability of the spatial metaphor 'scale' to access a large semantic document space. An experiment was conducted to examine whether different user groups can associate graphical changes in resolution in spatialized views with changes of level of detail in an index hierarchy of a digital document collection. Test participants were asked to utilize zoom tools to explore a spatialized subset of the GeoRef database, an extensive collection of geology and earth sciences documents. The outcomes of the experiment suggest that people are able to associate graphical changes in resolution of spatialized views (zooms) with changes in levels of detail of a document collection (hierarchical order). These results are independent of user group membership, but for some displays it takes people longer to make a decision.

Keywords. Spatialization, scale, usability, semantic spaces, spatial metaphors

1 Introduction

Research in information visualization deals with the design and implementation of computer-supported, interactive, visual representations to amplify peoples' cognition (Card et al., 1999). Such kinds of 'visual knowledge tools' (Card et al., 1999) are also known as spatializations, as they are very often based on spatial metaphors. Spatializations can be combined with information access techniques to help people find relevant pieces of information buried in large data archives. Distance (similarity between data items), regions (aggregation of similar data items), and scale (level of detail in a database) are examples of spatial metaphors utilized in information visualization. Although a large and diverse set of visual forms has been produced in information visualization, only recently have researchers in this young research field recognized the importance of empirical evaluation of their products. One of the main challenges in information visualization is the identification and establishment of a solid theoretical framework (Catarci, 2000), allowing the derivation of sound formalisms for successful information visualization designs, and effective graphical user interface implementations. One way to build solid foundations for information visualization is by providing empirical evidence of its usability. According to Robertson (1998), a survey of the IEEE symposia on Information Visualization between 1995-1997 revealed that only 6% of the papers presented dealt with information visualization usability. This has not improved substantially today, with the notable exception of a few, very recent studies reviewed in two special journal issues in the HCI and Information Science communities. One issue is devoted to individual differences in

Published as: Fabrikant, S.I. (2001). Evaluating the Usability of the Scale Metaphor for Querying Semantic Information Spaces. In: *Spatial Information Theory: Foundations of Geographic Information Science. Conference on Spatial Information Theory (COSIT '01), Lecture Notes in Computer Science 2205,* Montello D. R. (ed.), Springer Verlag, Berlin, Germany: 156-171. Color plates of the appendix and Errata information is accessible at URL:
http://www.geog.ucsb.edu/~sara/html/research/cosit01/appendix/index.html .

virtual environments (Chen and Czerwinski, 2000), the other deals with empirical evaluation on information visualizations (Chen et al., 2000).

This paper aims at contributing to a solid theoretical framework of spatialization, thus on the use of spatial metaphors for information visualization. The theoretical scaffold is based on geography and GIScience (Fabrikant and Buttenfield, 2001). This study provides answers to the question if people are able to associate graphical change in resolution (zoom-in) with different levels of detail in a document collection (hierarchical order). Empirical evidence is provided on the usability of the scale metaphor for graphically exploring a large text data archive. In addition, design guidelines are presented as suggestions for improvement of future spatialization displays.

2 Background

2.1 The Scale Concept

Scale is one of the most fundamental yet poorly understood phenomena in research dealing with geographic information (Montello and Golledge, 1999: 3). Scale has multiple meanings that not only differ across disciplines (mathematics, cartography, psychology, etc.), but also within a subfield. For example, in GIScience scale can be understood as resolution (i.e. 30m sampling interval of a digital elevation model), as level of abstraction (i.e. a map at 1:200,000 scale), as point of observation (i.e. vista space), or as organizational strategy (i.e. semantic hierarchies). In geographical analysis, scale relates to the resolution of items under study, and the level of detail that may be applied. The nature of the inquiry and the phenomenon of interest set the scale, and scale in turn determines the degree of generalization. The continuum of geographic scale is one of the most fundamental characteristics of geospatial data. Scale operates within a logical geometric framework, or frame of reference. This reference frame is the higher order spatial concept of hierarchy. Hierarchy is composed of the spatial primitives identity, location, magnitude and connection (Golledge, 1995). In geography, a reference frame is based on a formalized coordinate system (e.g UTM coordinates). Once the construction details of a chosen frame of reference are established, scale change can be identified and measured.

Scale is not only fundamental to geographical analysis, but is also associated with cognitive and experiential properties of the real world. Hierarchical order is a basic organizational principle of human cognition. Hierarchy is an example of the part-whole cognitive image schema (Lakoff, 1987). Empirical evidence shows that hierarchical order is an important aspect of how humans learn about the environment (Golledge, 1999). Human cognition varies with scale, ranging from personal-scale space with direct sensory interactions, to larger-scale space, where direct sensory interaction might not be feasible. Different dominant and subordinate levels of detail are evident and stored in a human's cognitive map (McNamara et al., 1989).

2.2 Scale as Metaphor

Scale is also a key element of research within the human-computer interaction (HCI) community to deal with the information overload problem. As volumes of collected data are increasing at exponential rates, identifying successful designs for dynamic, scale dependent information displays have become one of the most important factors for successful information visualization (Card et al. 1999). Because hierarchies, such as directory and file systems, are amongst the most common organizational structures in computing, it is not surprising that they are also one of the earliest and most popular examples for spatialization. An entire chapter in Card et al. (1999) is devoted to the scale issue, under the heading of 'Focus and Context'. Unfortunately, the theoretical basis for 'Focus and Context', its implications on design solutions, and the empirical evidence on the effectiveness of such techniques for information access are missing in this review. There seems to be little empirical research that reports on the usability of the scale metaphor in information visualization. Two recent studies deal with evaluating space-filling hierarchical structures (Pirolli et al, 2000; Stasko et al., 2000), but only Pirolli et al.'s (2000) work focuses directly on users' cognition and attention of the underlying spatial metaphor for visual information seeking. This investigation seems also unique in that it is placed within the context of a coherent theoretical framework borrowed from biology, that is information foraging theory.

As the creation of a semantic space is based on a metaphorical mapping from physical space (source domain) into a conceptual space (target domain) it is necessary to identify the appropriate source and target domains for usable spatializations. As shown above, scale is a good candidate for this metaphorical mapping (e.g. semantic transformation), as it includes a rich array of useful submetaphors (e.g. physical, cognitive, experiential etc.). The metaphorical mapping operates on a semantic and semiotic level. Cartographic design principles and geovisualization approaches offer sound semiotic foundations to depict the semantic mapping adequately.

Scale as understood in Geography and GIScience (source domain) is metaphorically mapped to the granularity of a text data base entity (target domain). The database in question is GeoRef, an online catalog of references for literature in the geological sciences. Scale change is metaphorically mapped to the change of level of detail in a document collection. Using GeoRef's indexing scheme as frame of reference, a semantic hierarchy can be constructed consisting of a single document at the finest granularity (e.g. depicted as a point symbol), a group of documents based on the same subject heading on a higher level (e.g. point clusters, identified by color), a theme containing multiple subject headings on a next higher level (e.g. a collection surface with colored regions), and so on until the highest level of detail is reached which consists of the entire document collection itself. In GeoRef, index terms are organized hierarchically in up to three levels (Goodman, 1997). For a given term many relationships can exist in an index, such as broader terms (a group of which a given term is a member), narrower terms (specific members of a group), related terms (synonym or alternative term) or geographic terms (location on the globe). Point symbols, point clusters and surfaces are the semiotic target domains for this study. Extending the surface metaphor into a third dimension provides opportunities for capturing the cognitive image schema of 'more is up' and 'near-far'. Documents that are densely clustered tend to metaphorically 'pile up' into semantic mountains. Using the scale change metaphor, graphic zooms allow users to navigate from one level in the hierarchy to the next, thus changing the semantic level of detail at which they view the collection. When examining the scale concept, the question is how well users associate hierarchical order within the document collection with the graphic (spatial) level of detail shown in a spatialized view, represented by the geometric primitives point and surface.

3 Choice of Methodology

The consequence of a missing theory in information visualization is the lack of a generally accepted methodology to assess and ensure its usability (Chen and Yue, 2000). Most empirical work is based on the usability engineering paradigm, particularly on task-centered user interface design evaluation (Norman, 1993; Nielsen, 1993). Although usability engineering partly borrows empirical principles and approaches from cognitive psychology (Lewis, 1991), its goals are of much more pragmatic nature. The focus is on users and their success in fulfilling particular computer-based tasks through a graphic display (Morse et al., 2000). The relationship of users individual differences, such as their cognitive abilities, their socio-demographic profile, or their individual knowledge base (e.g. background and training) is often not systematically assessed. The lack of a pre-existing theory made a more exploratory, qualitative approach necessary, where direct observations (think-aloud protocols) were combined with descriptive statistical measures. Inferential statistical procedures were applied as well. Due to the small sample of test participants (n=12), results from confirmatory statistical procedures need to be used with caution and are only of suggestive nature. A post-test questionnaire was also administered. This test included the collection of quantitative user background data, qualitative user preference information, as well as a cognitive factor test on spatial visualization ability (Ekstrom et al. 1976).

4 Experiment

4.1 Testing Environment and Equipment

To preserve the real world context in which a spatialized interface might be used, all testing procedures involving graphic displays were carried out digitally. A research lab, equipped with several NT workstations, located at the Geography Department of the University of Colorado at Boulder, was used as the test site. All tests were designed for a 17-inch color monitor and required a pointer device (trackball) as input. Participant responses to closed-end questions were recorded electronically. Participants' think-aloud comments were observed and noted by the examiner. A paper and pencil questionnaire was also collected and later summarized.

4.2 Data and Test Material

In addition to the zoom tools, which serve as an example of a spatialized query metaphor, two-dimensional and three-dimensional spatialized views were utilized for the experiment. A subset of the GeoRef database including one hundred documents were extracted from a GeoRef CD-ROM (1998), based on a random assignment of keywords. Spatializations were generated from this subset of documents. The construction details go beyond the scope of this paper, but can be examined in Fabrikant (2000). The size of the 2D and 3D objects were within a table top space, both depicted on a 2D computer screen. A subset of the utilized spatializations is shown in the Appendix (2D screens only). The 3D displays could be explored interactively in all three dimensions. As the user interaction mode varies considerably between two and three dimensions, it was concluded that comparisons of test results between dimensions allow only tentative suggestions for further research. A separate study was concerned with identifying the appropriate display parameters for the spatialized views, and the choice of an appropriate input device (e.g. trackball). These aspects of the study will not be further discussed here. Interested readers should consult Fabrikant (2000) for additional detail. The testing procedure was implemented using the ArcView GIS. ArcView's scripting language Avenue was utilized to digitally record user interactions and responses through logging scripts.

4.3 Participants

Twelve test participants with a diverse set of backgrounds were selected randomly for the study. The stratified sampling method included four main groups for this experiment: Librarians, GIScientists, and Students with mixed backgrounds from the University of Colorado-Boulder, as well as people from the Boulder community. GIScientists are participants in the strata of domain professionals and have extensive training in spatial information handling (i.e. earth sciences, geosciences, GIS, cartography etc.). GIScientists were recruted from diverse research labs on campus (but not from the geography department) and from the private sector. Domain experts typically access GeoRef's collection on a regular basis. Non domain professionals are participants who do not necessarily have a geoscience background, but have high expertise in information access and dissemination (e.g. information scientists, librarians etc.). These people might help others to access online collections such as GeoRef, and/or are trained to maintain such digital collections. This participant group consisted of librarians from the Earth Sciences Library at the University of Colorado-Boulder. Students and staff from the geography department at the University of Colorado at Boulder are participants with a mix of topical backgrounds and expertise who represent the continuum between domain and non-domain professionals.

4.4 Test Design

The goal of this research is to identify the usability of the spatial metaphor 'scale change' to query a text document collection. The research question to be answered is "Can people associate graphical change in resolution (zoom-in) with different levels of detail in a document collection (hierarchical order)?"

A proto-typical query task was defined to test if the spatialized query tool captures the metaphor mapping appropriately. The following information access scenario puts the experiment in context. Imagine an employee of a risk management firm is looking for a variety of documents related to Earthquakes. In Shneiderman's (1998) query task action taxonomy this very well defined information need relates to "specific fact finding". A patron could identify the earthquakes subject heading in a card catalog or a digital catalog in a traditional library. The question is how this patron would proceed in the search if the collection did not contain earthquakes as a subject heading, but seismology instead? Different test questions were utilized that triggered participants to use a zoom tool to change the level of detail presented in order to find a very specific item in the database.

A two-by-two, within-subject factorial design was employed to assess the query task in a controlled fashion. The independent variables were categorized on data type (Factor I) and dimensionality (Factor II). Factor I categorizes two dimensional and three-dimensional representations. Factor II, data type, includes discrete (point) representations and continuous (surface) representations. The dependent variables are accuracy of response, response time, response time to first zoom, and type of zoom. Mean response time to first zoom indicates how long it takes test participants to use the zoom tools for the first time. Data were collected to evaluate if participants applied the tool correctly. For example, if a term is presented in the test question that is on a lower level in the index hierarchy than the themes presented in the current display, the zoom-in tool leads to the correct answer. Accuracy of response is not a crucial evaluation instrument, as the use of the appropriate zoom tool will automatically lead to the correct answer. Direct observation protocols collected during think-aloud test sessions complemented the quantified variables. Responses and response times were recorded digitally during test sessions.

Following related work (Kraak, 1988) it was decided to separately analyze and evaluate test results for two and three dimensions. There seems to be very little knowledge regarding appropriate subject testing procedures comparing two and true three-dimensional representations. Most of the literature in this domain reports on empirical work regarding static 2.5D, where the third dimension is graphically simulated through depth cues such as shadows and hidden lines. In some tests reported here, however, users had the option to rotate a 3D object in all directions before responding to the test question. It may be important to emphasize that the size of the 3D objects were within a table top space, depicted on a 2D computer screen. Nevertheless, as the user interaction mode varies considerably between two and three dimensions, it was concluded that comparisons of test results between dimensions would have to be done very carefully, thus allowing only tentative suggestions for further research. Due to software limitations, the 2D tests and 3D tests were not always identical.

4.5 Test Procedure

First, participants signed a consent form and were given a ten dollar recompensation for participation. A warm-up phase of five practice trials followed, which allowed participants to get comfortable with the testing environment. The main experiment started with a short on-screen introduction to the test, and provided participants with the necessary context and background information to get through the experiment.

The main experiment started with either a point representation, or a surface representation in 2D. Presented with either start display participants were asked to look for a document on a particular level of detail hidden in the document collection. Starting with a point representation in 2D also required using a pan-tool before giving a response. The presentation order was randomized using a balanced Latin Square design to avoid potential learning bias. The same procedure was followed for the 3D representations. Due to software limitations, and differing interaction modes, the 2D tests and 3D tests were not identical. It is also important to note here that the design for the test starting with points and the test starting with a surface were not identical. A zoom into a point starting configuration resulted in the same point configuration with a change in the spatial footprint (less area visible). The change in level of detail was simulated with larger points (close up) and more detailed labeling. The zoom-in for the surface display resulted in a point display. Change in level of detail is simulated with a change in data type. Test displays for the 2D tests are shown in the Appendix. Two different questions were shown, depending if the starting display was showing points or a surface. If points were shown first, the test question was "Which blue patch most likely contains documents that relate to Geomorphology?". For surfaces, the wording of the test question was "Which of the topics most likely contains documents relating to the San Francisco Earthquakes?". Additionally, a text box was displayed next to the zoom tools, to inform people that the tools should be used to help make their response. This information panel changed depending on the data type shown in the starting display. For points, it read "Use zoom and pan buttons to help make your response". For surfaces, the text read "Click on the green labels in map or use zoom buttons to help make your response". The starting displays containing a surface also contained labeled buttons on the screen (see Appendix for more detail). If the button with the correct label was clicked (labeled Seismology), it zoomed to the same display, as if the zoom in tool on the answer dialog box was used. If the wrong labels were clicked, a dialog box appeared containing an error message indicating that this category did not contain the desired document and to try again. The question and answer window contained three tools, a pan, as well as a zoom-in and zoom-out button. Zoom tools and pan buttons were enabled or disabled depending on the starting display. The tools were controlled by the scripting language, to dynamically adapt to user reactions. The *Next* button and the radio buttons for recording the answer were disabled until a zoom tool (or label button) was used. The *Next* button was only active if a response had been recorded.

4.6 Methods of Analysis

Analysis of Variance (ANOVA) was applied to compare mean response times and mean times to first zoom. Analysis of Variance is based on the assumption that the data are normally distributed. As the collected response time data failed the test for a normal distribution, a natural log transformation was performed. Log-linear models were utilized for investigating the relationships of the categorical variable zoom type (three levels: zoom in, zoom out, no zoom) with data type (surfaces and points) and dimensionality (2D and 3D). The Logit Analysis is based on the calculation of expected odds of category membership of a dependent variable, as a function of independent variables. An odds is the ratio between the cell frequency of belonging to one category of the dependent variable and the frequency of not being in that category (Knoke and Burke, 1980). The response variable for the tests containing the three categories “zoom in” and “zoom out” “no zoom” was used as the dependent variable. Independent variables included dimension and data type. Due to the exploratory nature of this research (e.g. missing hypotheses) an inductive analysis approach was adopted. The goal was to find a model that not only fits the data well, but is also as parsimonious as possible, to be easily interpretable. For this reason a hierarchical stepwise Log-linear model selection procedure was carried out first. The search for the most parsimonious model started with a fully saturated model that included all possible effects, followed by a stepwise elimination process where effects were discarded systematically to determine their significance to the model. Once the best model had been constructed the strength and direction of the relationships of the remaining effects were determined by estimating the model parameters. Finally, computation of the Logit allows to predict the likelihood of test outcomes based on the identified model parameters.

4.7 Results and Analysis

The result section is organized as follows: summary tables containing descriptive statistics are followed by descriptive graphs. Response time is discussed first, followed by outcomes from inferential statistics. ANOVA is applied on response times and Log-linear modeling on zoom types. The result section is followed by a discussion of relevant individual difference data, and where appropriate, complemented by anecdotal information that was extracted from think-aloud protocols. Table 2 gives an overview of the response times results for scale change metaphor.

Table 1. Summary statistics of response times for the scale change metaphor

2D

mRT	zoom	stdRT	answer	n	used*	corr.	%corr.	factor
89.67**!**	30.00	30.22	A	12	12	9	75.00	monochrome points
84.58**!**	11.08	12.07	B	12	12	11	91.67	colored surface

3D

mRT	zoom	stdRT	answer	n	used*	corr.	%corr.	factor
13.97	31.42	13.70	B	12	12	11	91.67	monochrome points
31.68	22.42	29.20	C	12	12	12	100.00	colored surface

mRT	Mean response time in seconds	n	Number of responses
zoom	Mean response time to first zoom	used	Number of responses retained*
stdRT	Standard Deviation of RT in sec.	corr.	Number of correct answers
answer	Correct Answer	%corr.	Percent of correct answers

* if instructions were not followed, the response was removed from the sample.

Figure 1 reports on mean response times until a response was given. It took participants almost twice as long in two dimensions to respond, when points were shown first, than if a surface was presented first. In 3D, for initial surface displays the response time was twice as fast as for point representations. The means for 2D and 3D starting displays with points are about the same. The standard errors of the mean are also quite similar for 2D and 3D point displays. The standard error is very high for both dimensions for the test with the longer response time mean. (**!** Please refer to URL below for corrected information: `http://www.geog.ucsb.edu/~sara/html/research/cosit01/appendix/index.html`)

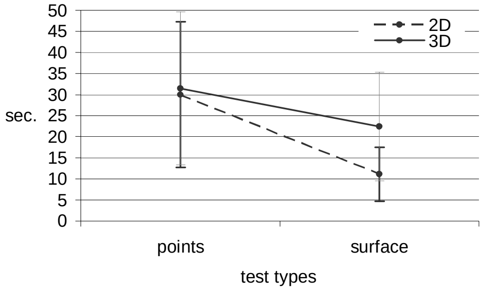


Figure 1. Mean response time for the scale change metaphor

Dispersion in response times is plotted in Figure 2. The spread is similar across groups. All but the student group seem to have outliers that increase the mean. Outliers are represented with a black dot in the box plot.

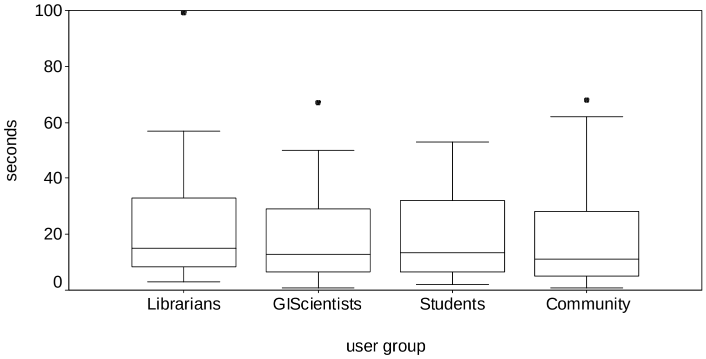


Figure 2. Variability of response time for the scale change metaphor

Figure 3 shows the amount and type of zooms performed for each dimension. The amount of zooms for points in 3D is substantially higher, and surfaces triggered much fewer zooms overall. There are slightly more zoom-ins than zoom-outs for points in both dimensions, which might lead to the conclusion that this test did not necessarily require participants to perform as many "control" zoom-outs to make sure that the right label was selected when responding. The number of zoom-ins is the same as for zoom-outs for surfaces in 2D and 3D. The amount of "control" zoom-outs, to match the found theme with the correct response label (toggle between views), increased when the level of detail changed through modifications in data type.

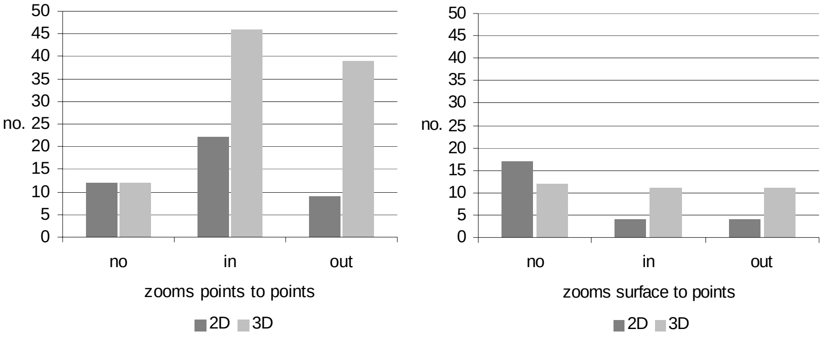


Figure 3. Magnitude of zooms across dimensions

This seems not surprising. Recalling a theme location and a label location from memory across displays is easier when the data type of the representation is kept constant. Moreover, the 3D test versions were not identical to their 2D representatives. In 3D, participants could not only zoom, but also rotate the displays in all directions. When participants chose to change the orientation of the starting configuration, the task of matching the spatial location of a theme after zooming in with its label in the start representation became even more difficult. It took participants much longer to respond. Considering this, it is surprising how well participants performed in the 3D surface case, where not only potential interaction could have made the task more difficult, but also the change of data type between displays added another level of visual processing.

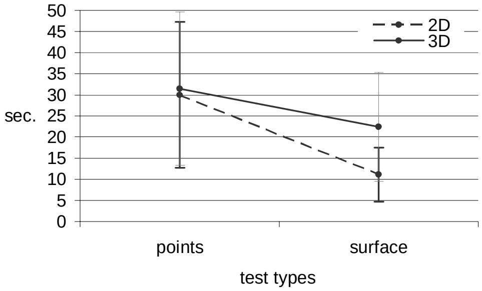


Figure 4. Response time to first zoom

Figure 4 plots the average time it took participants to first use the zoom tool. Starting with a point display it took participants much longer to respond for 2D and 3D. Librarians and GIScientists show a very similar pattern when breaking the time to zoom into participant groups (Figure 5). Zoom time for points is always higher across dimensions. For the student population exactly the opposite is the case, surfaces always trigger a longer response time to use the zoom tool across dimensions. People from the community have a similar pattern to Librarians and GIScientists in 2D, but show a similar pattern to students in 3D.

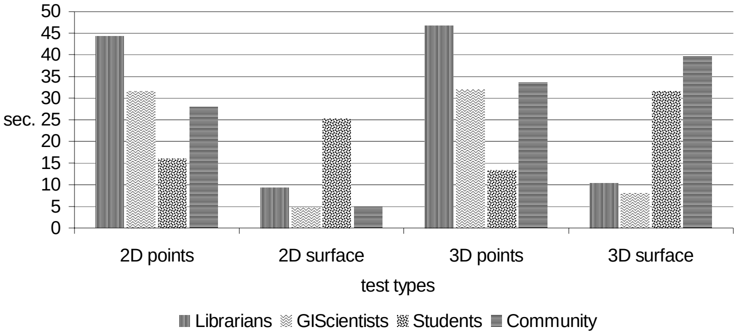


Figure 5. Response time to first zoom for dimensions, data type and groups

The ANOVA for mean response time to respond revealed a significant result for the main effect data type (F=5.29, $p < 0.05$) and the interaction effects for dimension and type (F=14.35, $p < 0.05$). It took longer to respond for points in 2D, than for surfaces in 2D. Significance for data type could include the possible interaction effect for the change in data type representation when people zoomed into surfaces. The ANOVA on mean response times and user groups did not reveal a significant relationship, nor did the ANOVA on the amount of zooms across groups. The ANOVA on zoom-ins, dimensions and data type yielded a significant main effect for data type (F=18.892, $p < 0.05$) and dimension (F= 6.95, $p < 0.05$). The number of zoom-ins is much higher for points than for surfaces. The ANOVA on the amount of zoom-outs not only resulted in significant main effects for data type (F=12.13, $p < 0.05$) and dimensions (F=12.13, $p < 0.05$), but also in an interaction effect (F=6.19, $p < 0.05$). 3D points trigger four times as many zooms than 2D points.

The analysis on zoom types, such as zoom-in, out and no zoom was performed with a Chi^2 test and Log-linear modeling. The Chi^2 analysis did not yield a significant result on the type of zooms performed across groups. For this experiment, the saturated Log-linear model that perfectly describes the collected data contains the nominal variable zoom type (Z), dimension (D), and data type (T) and is written with Lambda parameters as follows:

$$Z_i D_j T_k = \theta + \lambda_i^Z + \lambda_j^D + \lambda_k^T + \lambda_{ij}^{ZD} + \lambda_{ik}^{ZT} + \lambda_{jk}^{DT} + \lambda_{ijk}^{ZDT} \qquad (1)$$

where:

main effects

θ = effect of total factors in the scale test

λ_i^Z = participant zoom type repsonse

λ_j^D = dimension

λ_k^T = data type

first-order interactions

λ_{ij}^{ZD} = zoom response and dimension

λ_{ik}^{ZT} = zoom response and data type

λ_{jk}^{DT} = dimension and data type

λ_{ijk}^{ZDT} = zoom response, dimension and data type

In short hand notation the model can be written in terms of its highest order interaction only, that implies all lower order interaction effect. The hierarchical four-variable log-linear model is then written as follows (this notation will be used from now on):

$$G_{ijk} = \{ZDT\} \qquad (2)$$

Table 3 shows the list of possible hierarchical log linear models and associated changes in maximum likelihood ratios (Delta Lambda^2). Removal of model effects that result in significant Lambda^2 changes (e.g. $p > 0.05$) indicate that these effects are relevant to the model and should be retained.

Table 3: Hierarchical Log linear Models for the Scale Test

Backward Elimination (p = .050) for DESIGN with generating class ZOOM*DIM*TYPE

If Deleted Simple Effect is	DF	Δ L2	P	Iteration
ZOOM*DIMENSION*TYPE	2	0.7740	0.6792	3
DIM*TYPE	1	0.2280	0.6327	2
ZOOM*DIMENSION	2	15.2370	0.0005	2
ZOOM*TYPE	2	21.3340	0.0000	2

Shaded rows are selected for the Logit model. The best model has
generating class: {ZOOM*DIM}, {ZOOM*TYPE}
Likelihood ratio chi square = 1.00213 DF = 3 P = .801

The Logit model of choice for zoom types (Z) and graphic variables dimension (D) and data type (T) contains the following terms (see Table 3 for ΔL^2 values):

$$G_{ijk} = \{ZD\}\{ZT\} \qquad (3)$$

Completing the model with the estimated parameters for two examples, the probability of a zoom-in with 2D points as a starting configuration is 83%, whereas not zooming in with 2D point clouds as a starting configuration is 63%.

4.8 Individual Differences

These results are also interesting in light of users' background data. Spatial ability, as measured by the paper-folding test (Ekstrom et al., 1976), did not reveal a significant correlation with overall response time (Spearman's r =-0.23, p = 0.12). There seems to be a significant negative correlation of spatial ability with time to first zoom, although it could be considered weak (Spearman's r =-0.31, p = 0.03). Participants show high computer literacy. They all use computers on a daily basis (100%), and 50% of the participants use digital archives on a daily basis (e.g. access data on CD ROMs). Half the participants use online databases at least weekly (e.g. Web of Science, etc.). The majority (75%) has not had any formal training in cartography (mean = 0.5 yrs) and graphic design (mean = 1 yr.). About half of the subjects (50%) are not trained in computer graphics (mean = 1 yr.), or information retrieval (mean = 4.75 yrs). A considerable percentage of the participants use graphics (75%) and geographic data sets (40%) on a daily basis. There seems to be a weak positive correlation between spatial ability and GIS training (Spearman's r = .289, p = 0.05), but none between mathematics training (mean = 5 yrs.) and spatial ability. Neither GIS background, nor mathematics training seem to have a relationship with overall response time or reaction time to first zoom.

The post-test questionnaire also assessed participants' subjective usability ratings and satisfaction with zoom tools and graphic displays utilized during the experiments. Overall people reacted very positively to the spatialized displays and query tool. Most participants were intrigued by the displays and would use them again to query a document archive (83%). Half of the subjects found them somewhat unique, and 42% found them very unique. Most participants were at least somewhat attracted by the graphics (83%), and found them somewhat interesting and worth exploring (67%). Seventeen percent found them very attractive, and very interesting (33%). Half the participants mentioned the use of color as a reason why they found the displays attractive. About half related their interest in the displays to being able to see an overall, spatial structure in 3D, including a combination of labels and graphics. The ability to rotate the graphics in real-time and at will seem to have been a powerful experience for many participants. Think-aloud protocols and direct observations confirm the overall very positive experience participants seem to have had interacting with the displays. Most people showed reactions of surprise or disappointment when reaching the end of the digital portion of the experiment. Most felt they had barely started and would have preferred to go on (this was least 30 minutes after the start of the experiment). Manifestations of astonishment, surprise and intriguedness were often apparent, particularly when participants could directly manipulate the 3D displays. Comments like "this is cool", or "this if fun" were recorded with most participants. One participant, after having responded to the test question, got quiet while rotating the 3D representation in all directions for a very long time. This person tried to identify a pattern in the 3D point cloud. Other participants zoomed in and out of displays just to track how the change in data type affected their conceptualization of the information space. No signs of apparent frustration were detectable for any participant. One person mentioned in the post-test questionnaire having been frustrated at times, because of "not knowing the interrelation between displays".

4.9 Discussion and Design Recommendations

Results from testing scale change suggests that people are able to associate graphical change in resolution (zoom-in) with different levels of detail in a document collection A statistically relevant association exists between zooming behavior and graphic representation. Zooms from point into point representations lead to the desired document faster than zooms performed from a surface display into a point representation. User group did not affect the use of the metaphor. Although results on reaction time until the zoom tool is applied are not significant, main effect type ($F = 8.22$, $p= 0.007$) and interaction of type and group come close ($F = 4.94$, $p= 0.006$).

The main effects data type and dimension (controlled variables) are associated significantly with types of zooms. Graphic variables utilized to represent the database graphically will influence the usage pattern of the scale change metaphor embedded in the query tool. The factors dimension and data type both were significant, either for response times or for type of zooms. These results are difficult to interpret. One problem is the amount of factors that needs to be controlled for the scale change metaphor to strengthen the result of the experiment. Still, as a starting point for further research these outcomes are important to consider. In related experiments on the spatial metaphors *distance* and *arrangement* the graphic variables employed to render the semantic spaces were also shown to be important modifiers. Adherence to cartographic design principles (e.g. "more is darker") enhances understanding of the metaphors. Color and shape are particularly strong visual variables in this study, as revealed in participants' responses to open-ended questions in the post-test questionnaire. This has direct consequences for the design of the spatialized views.

For query tasks that require zooming to change the level of detail of a document collection the primary design consideration relates to modifying the display from points to points or surface to points. Surface-to-point zooms in 2D, and point-to-point zooms in 3D to seem to work best when access time to find relevant documents is important. If the amount of zooms needs to be minimized, then surface-to-point representations seem to perform better overall.

5 Conclusions and Outlook

A usability evaluation was applied to a spatialized query metaphor, the zoom-in and zoom-out tool, to access a spatialized portion of GeoRef, a collection of geology and earth sciences documents. Response times and accuracy of responses of participants using the zoom tools were collected during experiments on querying spatialized views. A qualitative investigation was also pursued with the think-aloud method.

Results indicate that people are able to associate hierarchical document relationships in a collection with the spatialized metaphor of scale change (zooms) for the information access scenarios described. For some displays it takes longer to make a decision. An important finding is that analysis of group membership did not yield a significant effect. Regardless of participants' backgrounds, the tested metaphor seems to yield similar responses. The scale metaphor provides many research threads worth exploring. The post-test questionnaire revealed that display attractiveness was directly connected with the ability to manipulate graphic representations in 3D, thus being able to explore an object from different view points. Current GIS typically represent geographic space (large-scale space) as pictorial space, or map space that are both non-manipulable space types (Freundschuh and Egenhofer, 1997). Assuming that direct manipulation and interaction with spatial representations enhances human's cognition of spatial primitives and spatial object relations, could this "empowerment" factor improve the usability current GISystems? Direct manipulation interfaces with iconic representation of system commands for spatial analyses are not new, but these interfaces often lack validation through empirical usability evaluation. The interdisciplinary design framework adopted for this investigation, and derived design recommendations based on empirical results might be a starting point to reconsider the construction and design of more intuitive iconic GIS interfaces?

Another research arena relates to "scalability" of the represented space. Human conception of space and spatial behavior are experience-based and scale-dependent (Freundschuh and Egenhofer, 1997). This study examines how people respond to a spatialized query tool to explore spatializations within a manipulable object space. This begs the question of how people's association with scale change would differ, if the object space changed in scale? How would these spatializations be understood within a virtual semantic document space? Information seekers could navigate in full immersion through a traditional library, and could also manipulate and interact with more abstract spatialized representations to search for information. What is the optimal balance between increased realism for intuitiveness and wayfinding, and the level of abstraction to reduce cognitive overload? How would wayfinding and navigation in such large-scale virtual spaces affect spatial metaphor comprehension? Scale change in object space has implications on empirical testing procedures. It is doubtful that current usability evaluation methods are adequate for virtual environments (VE). These and other challenges have to be addressed, to maximize the potential that spatialization has to offer for knowledge discovery.

6 Acknowledgements

This study is a subset of the author's dissertation completed at the University of Colorado at Boulder. Thanks are due to my Ph.D. advisor Barbara P. Buttenfield for her support and guidance throughout this study.

7 References

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in Information Visualization. Using Vision to Think,* Morgan Kaufmann, San Francisco, CA.

Catarci, T. (2000). What's New in Visual Query Systems? *Proceedings,* First International Conference on Geographic Information Science, Savannah, GA, Oct. 28-31, 2000. (http://www.giscience.org/GIScience2000/invited/Catarci.pdf)

Chen, C., Czerwinski, M., (Eds.) (2000). Special Issue on Empirical Evaluation of Information Visualisations. *International Journal of Human Computer Studies,* 53(5).

Chen, C., Czerwinski, M., and Macredie R. (Eds.) (2000). Special Issue on Individual Differences in Virtual Environments. *Journal of the American Society of Information Science,* 51(6).

Chen, C. and Yue, Y. (2000). Empirical Studies of Information Visualization: A Meta-Analysis. *International Journal of Human-Computer Studies,* 53: 851-866.

Ekstrom, R. B., French, J. W., Harman, H. H., and Derman, D. (1976). *Kit of Factor-Referenced Cognitive Tests.* Princeton, NJ, Educational Testing Service.

Fabrikant, S. I. and Buttenfield, B. P. (2001). Formalizing Semantic Spaces For Information Access. *Annals of the Association of American Geographers*, 91: 263-280.

Fabrikant, S. I. (2000). Spatial Metaphors for Browsing Large Data Archives. *Unpublished Disseration,* University of Colorado-Boulder, Department of Geography, Boulder, CO.

Freundschuh, S. M. and Egenhofer, M. J. (1997). Human Conceptions of Spaces: Implications for GIS. *Transactions in GIS,* 2: 361-374.

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

Goodman, B. A. (1997). *GeoRef Thesaurus.* Alexandria, VA, American Geological Institute.

Knoke, D. and Burke, P. J. (1980). Log-Linear Models. *Sage University Paper Number 20.* Newbury Park, CA, Sage Publications.

Kraak, M. J. (1988). Computer-Assited Cartographical Three-Dimensional Imaging Techniques. Dissertation. Delft, The Netherlands, Delft University Press.

Lakoff, G. (1987). *Women, Fire, And Dangerous Things: What Categories Reveal About The Mind.* Chicago, IL, University of Chicago Press.

Lewis, C. (1991). Inner and Outer Theory in HCI. In *Designing Interaction: Psychology at the Human-Computer Interaction Interface,* Carroll, J. M. (ed.), Cambridge, MA, Cambridge University Press: 154-161.

McNamara, T. P., Hardy, J. K., and Hirtle, S. C. (1989). Subjective Hierarchies in Spatial Memory. *Journal of Experimental Psychology. Learning, Memory and Cognition,* 15: 211-227.

Montello, D. R., Golledge, R. G. (1999). Scale and Detail in the Cognition of Geographic Information. *Varenius Project Report,* Specialist Meeting, May 14-16, 1998, Santa Barbara, CA.

Morse, E., Lewis, M., and Olsen, K. A. (2000). Evaluating Visualizations: Using a Taxonomic Guide. *International Journal of Human Computer Studies,* 53: 637-662.

Nielsen, J. (1993). *Usability Engineering.* Boston, MA, Academic Press.

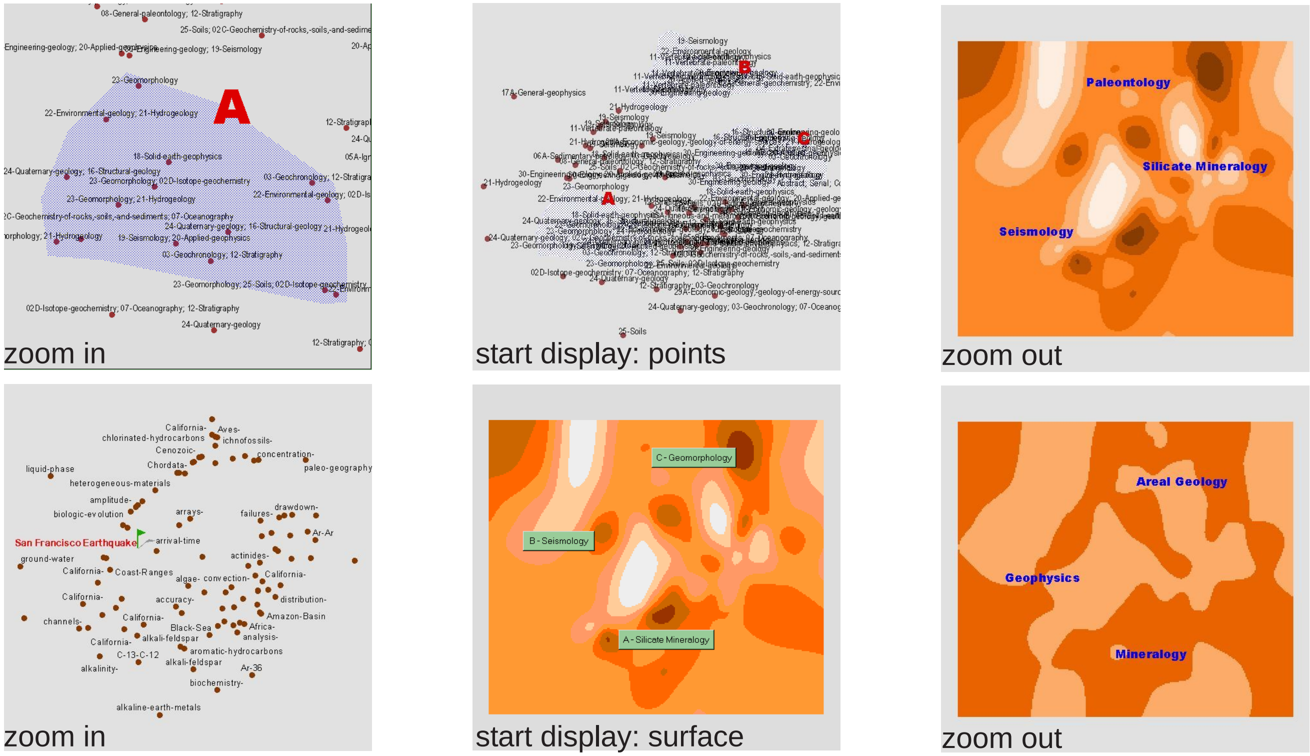
Norman, D. A. (1993). *Things That Make Us Smart. Defending Human Attributes in the Age of the Machine. Reading,* MA, Addison-Wesley.

Pirolli, P., Card, S. K., and Van Der Wege, M. M. (2000). Visual Information Foraging in a Focus + Context Visualization. Xerox PARC, Technical Report, IR-R-2000-14. URL: http://www.parc.xerox.com/istl/projects/uir/pubs/pdf/UIR-R-2000-14-Pirolli-TechReport-VIFinFocusContext.pdf (Jan. 2001).

Robertson, G. (1998). Keynote Address: Leveraging Human Capabilities in Information Perceptualizationd. IEEE Symposium on Information Visualization (InfoVis '98), Oct. 19-20, 1998, Research Triangle Park, North Carolina. URL: http://www.research.microsoft.com/~ggr/infovis98.ppt (Jan. 2001).

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Reading, MA, Addison-Wesley.

Stasko, J., Catrambone, R., Guzdial, M., and McDonald, K. (2000). An Evaluation of Space-Filling Information Visualizations for Depicting Hierarchical Structures. *International Journal of Human-Computer Studies,* 53: 663-694.



zoom in

start display: points

zoom out

zoom in

start display: surface

zoom out

Appendix. 2D test screens for the scale change metaphor

Note to the reader: The test screens for the 3D tests can be accessed at the following Web address: http://www.geog.ucsb.edu/~sara/html/research/cosit01/appendix/