Visual Classification with Information Visualization (Infoviz) for Digital Library Collections

Judith Gelernter*

School of Information and Library Studies, Rutgers University, 4 Huntington Street, New Brunswick, New Jersey 08901 USA, <gelern@scils.rutgers.edu>

Judith Gelernter earned her undergraduate degree from Yale University, her master of arts in fine arts from Harvard University and her master of science in information and library science from Simmons College. She is a doctoral candidate in information science at Rutgers University. Her research in interface design and usability for information systems mirrors the practical skills she uses at Knowledge Edge, the consulting company she founded in 2004.

* The author wishes to thank Michael Lesk, who discussed the essence and the boundary of this topic, Tefko Saracevic for his encouragement, and Chaomei Chen for his comments.

Gelernter, Judith. **Information Visualization (Infoviz) for Digital Library Collections in Digital Libraries.** *Knowledge Organization, 34(3)*, 128-143. 61 references.

ABSTRACT: The purpose of information visualization (infoviz) is to show information graphically. That purpose is often obscured by infoviz designs that are not well understood in practice. This paper offers an overview of infoviz culled from the literature on applications of information visualization for the digital library: how the clustering works that creates the topics, and how those topics are represented graphically. A taxonomy of infoviz designs in one, two and three dimensions is presented. It is suggested that user evaluations of infoviz designs might be used to enrich infoviz theory and, whether through application of the theory or through application of user remarks, developers might improve infoviz interface comprehensibility. Design recommendations are made in an effort to improve weaknesses and capitalize on strengths of present interfaces in representing knowledge visually.

1. Introduction

Information visualization (infoviz) is running away from interface design in aspiring innovations and aesthetics. Judelman (2004, 245) wrote, "InfoVis could tap into the ideas and experiments in the arts to help expand the repertoire of visualization strategies." But do we want creativity, or clarity? With this paper I give background on infoviz and its use in classification of digital collections, beginning with an overview of the mechanics of creating categories and plunging into recurring questions in the literature. Taxonomy of infoviz graphics is based on the use of one, two, or three dimensions of the computer monitor's surface. The present lack of a theory of information visualization might suggest why one design would be better than another. A sample of infoviz design variety is shown in this paper. Alternatively, user evaluation could provide evidence as to why one design should be used rather than another. In the words of Chen ([1999] 2004, 1), there is a present "lack of generic criteria to assess the value of information visualization, either independently, or in a wider context of user activities." Should infoviz interface design pertain more to the user group or to the collection content? Analyzing carefully what is now available in light of user preference and task performance could allow us to retain current strengths while improving upon weaknesses.

In essence, infoviz interfaces should aim for what works for users, especially clarity. We seek keys to information display, whether it is using traffic light red for "stop" and green for "go," or organizing information visually in other ways that are widely recognizable. Most digital libraries have no standard by which to organize information, as Dewey or Library of Congress classifications are used to organize books in libraries. Some online databases have no graphical overview at all, and users might be forced to rely on entering terms in a keyword search box to determine database contents. In fact, Fast and Sedig (2005, 8) remark that, at present, "digital libraries that rely on information visualization techniques are the exception, not the rule."

The rank variety of infoviz interfaces and lack of standardization suggests that the genesis of such interfaces is in its infancy, in that designs tend to simplify and stabilize over time. Think of the rotary dial interface for a telephone, for example, which was dominant from the 1930s through the 1980s. Present infoviz designs show, arguably, too much. They show the range or depth of information organization (whether faceted or hierarchical), the scope of organization on a single screen (broad categories such as those in the *Dewey Decimal Classification* or close such as those in the Library of Congress *Classification*), and how the user accesses those categories (interaction design). Designs tend to convey categories by way of a list, a 2D graphic, or a 3D landscape. Some designs use icons, but more often, they use morphograms–lines, circles and simple geometric figures.

A visualization is supposed to be a tool that relieves working memory for more important work, such as judging what information in the database to select. Tversky et al. (2007, 72) give the analogy of pen and paper as tools used to visualize numbers and help people do arithmetic. Poorly designed visualizations do the opposite because they take time and processing power for the user to understand.

Information visualization is over a decade old, and has been facilitated by the decreasing cost of storage and the increased availability of high resolution monitors. Discussions of visualization methods can be found in surveys of information visualization and interface design. More often they are published in journals of information science rather than computer science. Some schemes have been developed collaboratively by scientists in information retrieval, advanced graphics and automated analysis of algorithms; others have been developed by computer scientists. Papers about infoviz in digital libraries have been produced by research centers in South and North America, Europe, the Middle East, and Asia.

A highly specific infoviz vocabulary is being developed. Feng, Jeusfeld and Hoppenbrouwers (2005) refer to a knowledge subspace for representation, which is in contrast to a document subspace for data. A knowledge subspace elsewhere in the literature is called cartography (Polanco, Francois and Lamirel 2001), domain visualization (Chen 2000), or visual thesaurus (Ramsey et al. 1999). Representations based on a spatial metaphor are called spatialized views or information spaces (Fabrikant 2000, 65). Marshall and Madhusudan name knowledge elements of concept nodes, concept-link-concept propositions (also called cross-links), hierarchical clusters and other map substructures (Marshall and Madhusudan 2004). Concepts on the same level of hierarchy may be called neighbors or siblings; those on different levels, parent and child. The hierarchical arrangement of clusters is called nesting.

2. The Purpose of Visualization Schemes in Digital Libraries

Infoviz graphics have been used to display information abstractly, thus infoviz interfaces would seem ideal for digital libraries because it has been found that users often are unsure as to the range of a digital library's content, and some users complain they cannot understand a digital library (Saracevic 2004, 8). An earlier paper by Greene, Marchionini, Plaisant and Shneiderman (2000, 380) also remarks that digital libraries "regularly fail to provide honest representation of what they include, wasting users' time, and increasing their frustration with online systems."

Visualizations could help by showing contents at a glance. Information visualization, by allowing users to select from among content interactively, becomes a tool for browsing. Where keyword search is beset by ambiguity of term selection and vocabulary mismatch, selecting visual categories provides an "alternative means of assessing intellectual structures" (Chen 2000, 261). The broad categories offered on an initial screen are consonant with what Belkin, Oddy and Brooks (1982) posited as the user's Anomalous State of Knowledge (ASK) at the beginning of an information search. Theoretically, the user in the ASK state is uncertain as to what he seeks. Rather than rely on keywords to sift among specifics, therefore, the uncertain user of an infoviz interface could choose

among broad categories and then filter to specifics. This hypothetical usefulness of visualizations of categories to dispel some uncertainty at the beginning of the information-seeking process should be confirmed by scientific experiment. Börner, Chen and Boyack (2003) believe there will be a time when domain visualization will help answer actual questions. Similarly, Feng, Jeusfeld, and Hoppenbrouwers (2005) foresee a cognitive function for enhanced digital library systems in which users receive not only documents requested, but also intelligent answers to questions.

3. How Infoviz Works

How are the categories and the relationships among those categories determined? How do these categories then translate into a visual scheme? Infoviz applications require mathematics and programming for complex layouts and dynamic graphics. This section is an overview of what is involved. Those looking for in-depth discussion of how backbones are extracted from large data sets and the techniques for generating spatial layout might consult the second and third chapters of Chen's *Information Visualization: Beyond the Horizon*, 2nd edition, 2004.

3.1 Division into Topics

Contents divide into topics through clustering or classification. Clustering precedes classification and creates the categories into which items fit. In clustering, topics are assigned automatically. Clustering is considered a bottom-up approach because it goes from documents to categories, as opposed to classification, a top-down approach that starts with categories and then assigns documents to those categories. In classification, documents may be clustered manually or assigned to categories automatically. Sebastiani (2002) reviews several approaches to automatic categorization.

In the first step, an algorithm indexes documents and identifies keywords and patterns. After similarities among documents are determined by the indexing, another algorithm forms document clusters. The two predominant clustering techniques are agglomerative and partition-based. The agglomerative technique is prohibitive for large collections because it categorizes data points individually. Partition-based clustering techniques include K-means, Naïve Bayes, Gaussian mixture model, Latent Semantic Indexing, Pathfinder network scaling, and Kohonen self-organizing maps (SOMs) (Krowne and Halbert 2005). The drawback of automatic clustering, as illustrated by Hearst (1999), is that it may result in categories with different levels of specificity.

3.2 Representing the Topics

Ordering techniques such as triangulation or force directed placement generate a layout that shows relevancies among topics as distances among clusters. Topics may be alphabetized linearly, represented graphically in two-dimensional approaches such as the concept map, the tree map, and the Kohonen SOM, or spatially in three-dimensional approaches. 3D domain landscapes, like 2D maps, represent the relationship among topics by cluster proximity, and the number of documents per cluster by node size in two dimensions, or by node height in three dimensions. 3D approaches alternatively have been termed 2.5D visualizations (Boyack, Wylie & Davidson 2002a, 147). Distortion-based techniques allow a visual overview, expanding or shrinking relevancies as needed to fit within display boundaries.

An abstract scheme may be masked behind an interface assumed to be more intuitive. For example, Merkl and Rauber (2000, 436) describe the SOMLib project whose interface displays books on shelves. The size of a document is symbolized by book thickness, the time of last access is symbolized by proximity to the shelf edge, and the disuse of documents by the appearance of dust or cobwebs. Frequency of use is shown by a book looking well-thumbed, and newness is shown by a shiny book cover. Different document types appear on the shelf as different types of media.

3.3 Interaction Design

Shneiderman and Plaisant (2005, 581) simplify the mechanics thus: "Overview first, zoom and filter, then details on demand." Filtering and zooming are used to browse the information space and navigate through its content. The term used is "drill down," meaning to proceed from generalities to specifics.

4. Recurring Themes in the Literature

Insight into the infoviz design process is suggested by the recurring themes in the literature on infoviz in digital libraries. Here are some themes:

- What terms should be used for the categories, and how should those terms be displayed?
- Is there any way for the user to enter category terms?
- Should relationships among the documents be depicted by hyperlinked words or by graphics?
- What should be the distance between nodes for relationships on the same level of specificity?
- How many aggregate nodes should be displayed at each level?
- How should node relevance be depicted—with color, size, texture, shape? Should these attributes be adjustable through user options?
- How could navigation work so that when a user goes from a topic node to the next level of specificity, sight of the whole is retained? By means of animation and zoom? by means of multiple views or windows to represent simultaneously different levels of specificity? Or by access to browse history?
- How could the end of a hierarchy be depicted so that a user knows he has arrived at the transition between visualization and document?
- How should the interface be designed to allow querying of the result display?
- What visual metaphor could be adopted to make an interface recognizable immediately?

The variety of graphical and interaction design questions suggests that digital library visualizations are new and in flux, and there is as yet little standard practice. So as would be expected, a wide range of design options are included in a toolkit for developers of infoviz interfaces, call *Prefuse*. The *Prefuse* software provides developers with visual graphs, node-link diagrams, and ways to structure free data such as scatter plots and timelines (Heer, Card and Landay 2005). Seemingly absent from the literature are explanations of why one particular visualization has been chosen over another – the topic of this article.

5. A Sample of Information Visualization Schemes in the Literature of Digital Libraries

Visualizations may be organized in any number of ways: by space, with changes in data over time, by number of info-bearing attributes of the visualization (color, animation, shape, size), or by low level tasks that the visualization should facilitate (Amar and Stasko 2004, 145). The classification below is selected because this is how it typically appears in the literature. For example, Morse and Lewis's (2000) taxonomy of visualizations includes four types: word, icon, graph and physical analogue. Word-visualizations correspond to the 1D scheme described here, include hierarchies in a list or menu or site map, and are about the most common. Icon- and graph-visualizations are species of the 2D schemes shown here (although the concept map in this paper fits neither category). Concept maps, tree maps and SOMs are less common. Objects in the physical world such as mountains correspond to 3D schemes. 3D spatial representation has been called popular and controversial (Shneiderman and Plaisant 2005, 585), but its use so far in digital libraries is limited.

In the section below, examples are chosen to be representative but not comprehensive. Details of interaction design are not discussed because they represent an element distinct from the graphical scheme at issue here. Some schemes represented are intended for initial-browse interfaces and some for results display. Implications are discussed later. Each additional level of spatial representation affords additional ways to convey information about the data. A 1D list overview allows at-a-glance viewing of an amount of data too large to be comprehended individually. 2D representations can show the intersection among clusters, that is, topics. Some 2D representations can show the degree of relationship among categories, as is done with lesser precision in the "broader term," "related term," "narrower term" of a classification hierarchy. 3D schemes have the ability to show more characteristics of the collection itself.

5.1 Hierarchical List

Strictly speaking, the hierarchical list in alphabetical order (as in Figure 1 below) is not information visu-

alization because there is no representation of concepts graphically. The example is included here to show a range of possibilities in the spatial display of information. For a hierarchical list, categories are limited in order to accommodate a single screen so that the user does not need to scroll to view the whole at a glance. The list might be organized by topic, time period, by format, or by location, and all four alternatives are offered by the *American Memory* digital library of the Library of Congress. The excerpt below shows a topical list with subtopics.

Advertising
Early Advertising, more

African American History
Slave Narratives, more

Architecture, Landscape
Historic Buildings, more

Cities, Towns
New York City Films, more

Culture, Folklife
September 11, Dust Bowl, more

Environment, Conservation
Florida Everglades, more

Government, Law
Continental Congress, more

Immigration, American Expansion
Chinese in California, more

Figure 1. Hierarchical list: Excerpt from topics represented in the American Memory Digital Library of the Library of Congress

5.2 Concept Map

A concept map is a sort of visual outline, as one might sketch by hand on a chalk board, where nodes are concepts and links are inter-relationships. The concept map is a type of 2D visualization that was born in and has remained in the service of education (Novak and Cañas 2006), and can be used for learning (Marshall et al. 2003, 137), (Sumner et al. 2005). Novak (1988, 227) gives instructions for creating a concept map by telling his audience they might find it helpful to write concepts on cards so that they can move them around in creating the map. Novak remarks that concept maps are generally hierarchical, with the most general concept at screen top and more specific topics lower down (Novak and Cañas 2006). Educationally, the idea is that learning comprises assimilating new concepts into existing ones, such that the maps of in-depth learners include more concepts, relationships and branches than those of new learners (Phillips, Rajkumar & Shao 2005, 5).

Pictorially, the concept map resembles a chart with relationships among categories shown by straight lines (see Figure 2). Similar is the star map with the main concept at center. The *Inxight* software used for the "at a glance" browse section of the *National Science Digital Library* also assumes a star shape. The *National Science Digital Library* contains a section with concept maps (http://strandmaps.nsdl.org). The map of a field shows related concepts. Clicking on a concept retrieves a "pop up" enlargement with URLs of websites illustrating the topic.

For large quantities of documents, such mapping becomes more difficult. The *GetSmart* system uses concept maps for such quantities, and was deigned for use in the *National Science Digital Library*. *GetSmart* was developed jointly by the University of Arizona and Virginia Tech to assist student learning and may be used by educators to assess the students based on the maps they draw. It is built with a taxonomy of relationships from which the user selects (superset, subset, component, proximity, causality), represented as different colors on the topic map. If the user chooses, each concept can be associated with one or more URLs. At the time of writing, *GetSmart* could be downloaded from http://feathers.dlib.vt.edu/~rshen/ConceptMap/GetSmart.html.

5.3 Tree Map

The original use of the tree map was to view hard disk storage space and recognize larger files as candidates for deletion when the disk was filling. Pictorially, it represents the relative quantity of documents in a cluster by rectangle size, and the relationships among the documents by rectangle proximity. Shneiderman et al. (2000, 58) confesses that "trees represented as node-link diagrams are useful, but as they grow to include thousands of nodes and many levels, layout and navigation problems become serious." Many users required training in earlier tree map versions. Novices

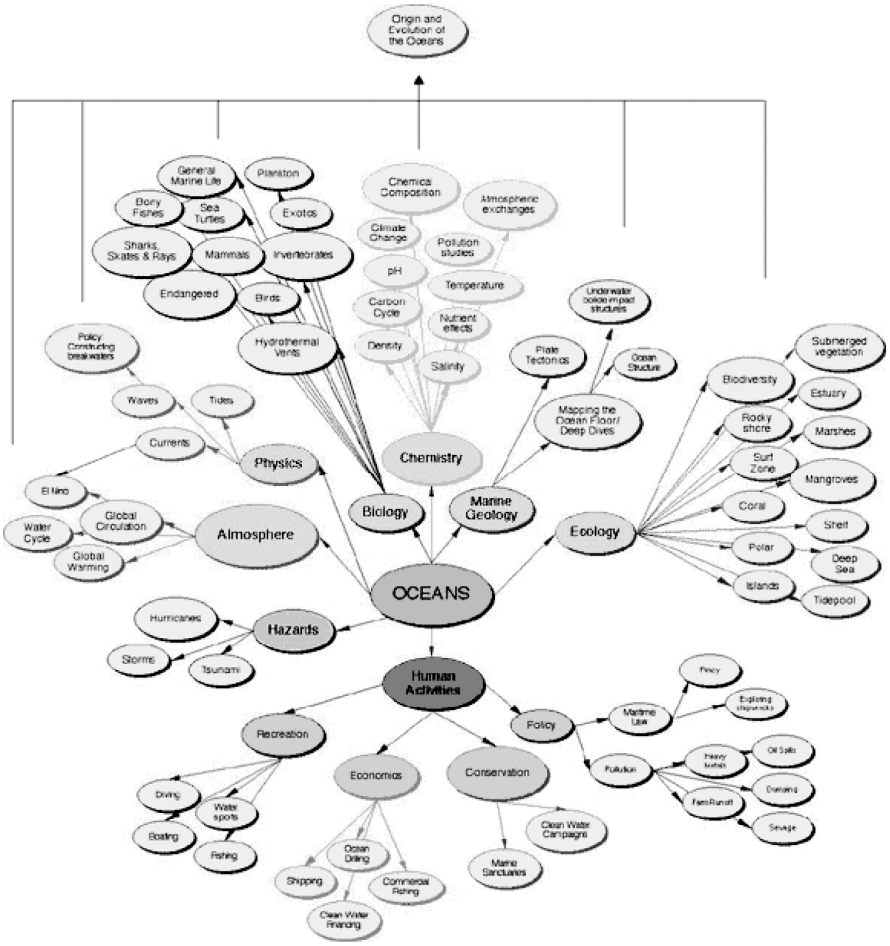

Figure 2. Concept map (http://www.csmate.colostate.edu/dwel/oceans_conceptmap.html)

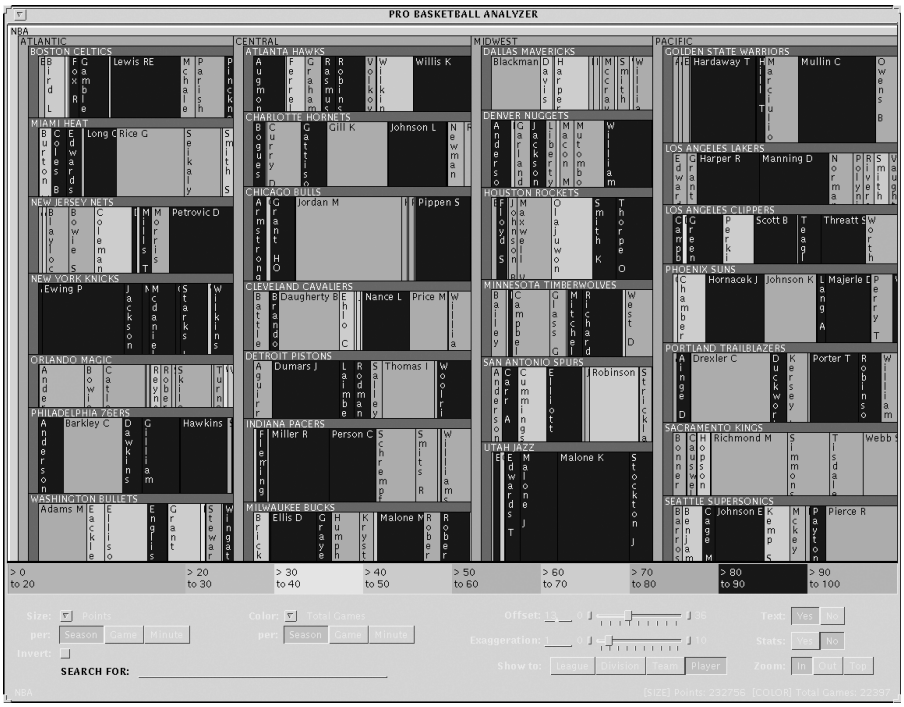

Figure 3. Tree map (http://www.cs.umd.edu/hcil/treemap-history/)

did better examining fewer nodes on fewer levels. Figure 3 is a sports application in which basketball players are organized into the teams that comprise the National Basketball Association Leagues, and size of player boxes is determined by points scored by fouls, free throws and such (Shneiderman 2006). A working tree map for the stock market is found in the map section of http://www.smartmoney.com/marketmap. The map uses size, value (color) and hierarchy to show real-time stock market data. Users may adjust the amount of labeling. A cursor over a section brings up cursory information, clicking over a section brings up a menu from which the user can learn about company news, earnings, ratings or other data. A *Tree Map Builder* available through the *SmartMoney* site provides the code base that can be licensed to suit any company's requirements.

5.4 Self-Organizing Map (SOM)

Kohonen developed the SOM to create a visual order for large quantities of information (Kohonen, 1990). An early, basic example is illustrated below. SOMs are a specific kind of Artificial Neural Network (ANN), a computer network modeled roughly after the neural structure of the brain. ANN's can describe nonlinear, non-obvious relationships between variables, so they work well for classification in the form of the self-organizing map. A SOM extracts features from documents which it represents as nodes, and the nodes adjust so that they will reflect similarities between clusters on the same level and on different levels of the hierarchy. Lastly the spaces are labeled. A cursor over a region might display the titles of documents most strongly associated with the region. The process is described, for example, in Hearst (1999), Lesk, ([1997] 2005), Deng, Zhang and Purvis (2004), and Goren-Bar and Kuflik (2005). A second layer of the map might show that category subdivided again, with the next layer either another subdivision of the topic or a results list.

Research as early as 1998 showed that a SOM could categorize a large, eclectic Internet data pool. Users who evaluated the system liked graphical elements of the map and the fact that the map had layers or levels (Chen, Houston, Sewell, Schatz, 1998).

Figure 4. Self-Organizing Map (Roussinov and Chen 1998)

The Austrian SOMLib Project uses growing neural network architecture (GHSOM) in which each layer consists of several independent SOMs (Merkl & Rauber, 2000). The SOMLib is masked by a book shelf graphic (as described in the section on how KDV Works) to make the interface more intuitive.

SOM can be used to manage huge information collections. As an alternative to Google-style keyword web search or the Yahoo directory, tests have been done to organize website home pages based on concepts found within them. In the example illustrated below, a first level SOM map shows the number and relationship among topics found in 10,000 entertainment-related home pages. Time to process the SOM varied according to input vector size (Roussinov and Chen 1998). Researchers used a collection of entertainment-related home pages extracted by a spider runner on the entertainment portion of the Yahoo directory (Chen, Schuffels, and Orwig 1996). Each region represents a topic determined by the SOM with a number indicating the number of home pages classified within each region. Click each region to find a more detailed SOM map representing that region's sub-topics; click on the lowest-level map to view the actual home pages. Problems arose from the fact that the areas of the maps are labeled, but the maps themselves are not, so it is not invariably clear to the user what level he is on. Also, in some cases map labels were longer than the map area.

5.5 Three-Dimensional representations

The construction of three-dimensional browse interfaces for digital libraries is called spatialization. One digital library called *CAVE-ETD* is an immersive simulation with aisles and shelves, like a physical library (Das Neves and Fox 2000). Domain landscapes have been created for subject and also for citation analysis (Chen 2000), (Börner, Chen, and Boyack 2003).

Figure 5 is a view from the SPIRE package from Pacific Northwest National Laboratories. Mountains represent theme clusters across documents, the proximity of the mountains indicates similarity of themes, their arrangement indicates concentration, and height of the mountain indicates the relative strength of the topic in the set. Other attributes that can be represented include individual document size, source, year of publication, semantic similarity, and document-query relevance (Chen 1999). *VxInsight* developed by Sandia National Laboratories in New Mexico also uses a landscape to show a knowledge domain (Boyack, Wylie, Davidson, 2002b).

Domain visualizations have been lauded for providing a good overview of complex document sets. The trained eye could use such an interface over a large set of, for example, medical documents, to see in what direction some type of research is heading, to find relationships among topics or research teams, or to identify key topics. But it has been noted that untrained users do not necessarily understand what they see, and if they do understand, they may make more unnecessary navigation turns to get to the documents than they might in a comparable system (Shneiderman and Plaisant 2005, 586).

6. Information visualization— Theory or Principles?

The previous section represents only a selection of infoviz interfaces to digital libraries that are found in the literature on digital libraries. Is one design somehow better, and if so, is it better in all circumstances, or only for some user groups or some types of digital content? Chen and Börner (2002) comment on the range of questions they found in the literature regarding representation schemes. Some questions are inseparable from particular environments, in their opinion, but more often, "the same fundamental question disguises itself in different forms" (229). Their fundamental question seems to be the same that underlies this paper: what is the best way to represent content graphically? To answer this question, we would need to know why one was better than another–that is, we would need some basis for evaluation. This answer would give a strong basis for a theory of information visualization. Chen and Börner admit that foundation works in the field are needed urgently, and that no principles from perception, cognition, graphics or computer-aided interaction readily lend themselves to design.

There is no theory of information visualization, per se. But even though a theory is lacking, researchers have proposed graphical design principles. One set of principles come from Edward Tufte, who teaches statistics and graphic design. Another set has been proposed by Shneiderman and Plaisant, who work in human-computer interaction and interface design. The two consider the same question from slightly different angles, and their viewpoints reflect their respective backgrounds and media. Tufte

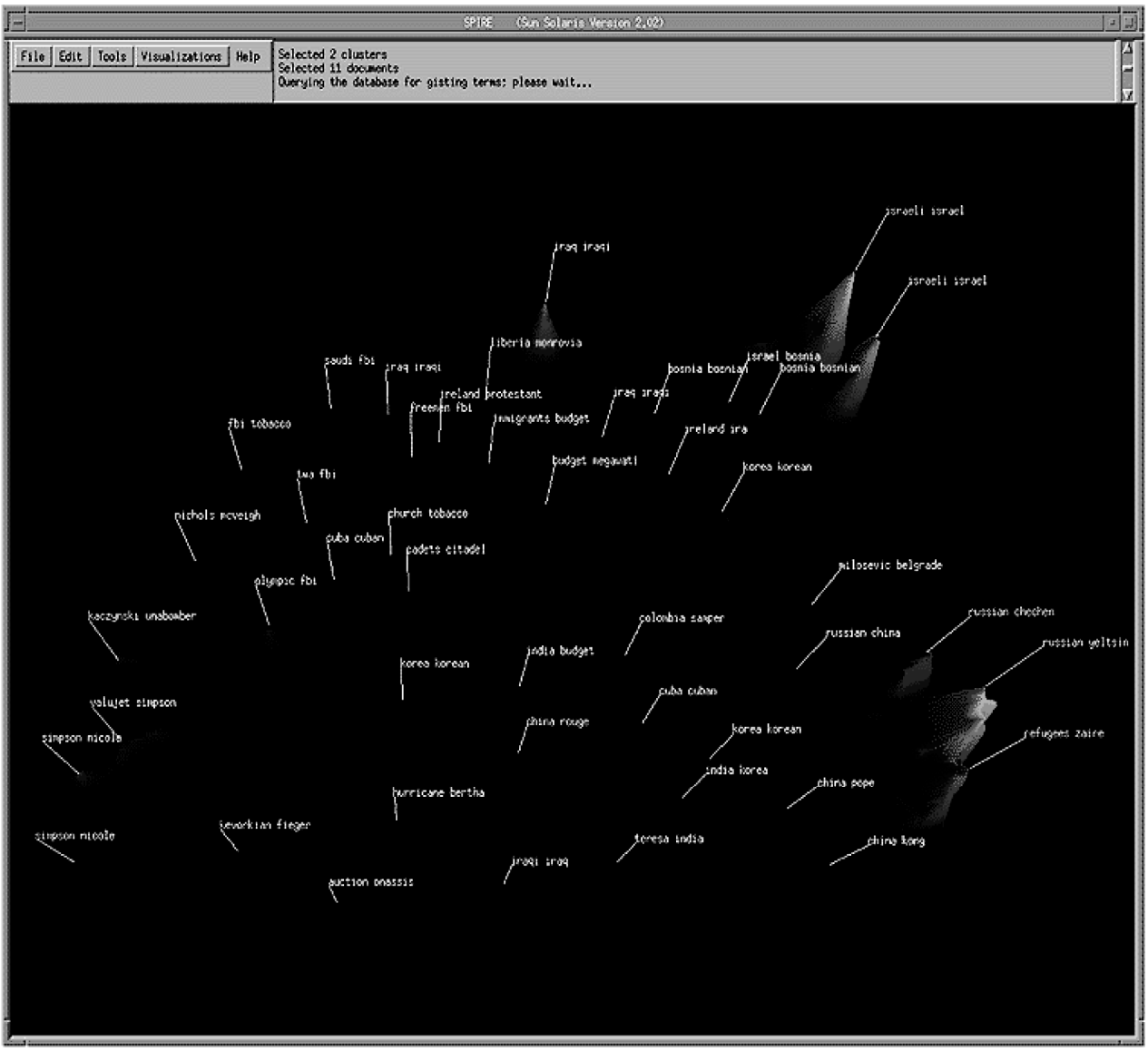

Figure 5. Domain landscape (Themeview from SPIRE visualization tool set, from Pacific Northwest National Laboratories. Courtesy of Dennis L. McQuerry)

([1983] 2001) is interested in the question of how to represent data clearly. He writes that graphic design demands three different skills: the substantive, the statistical and the artistic, and that the aim of graphic designers should be the clear portrayal of complexity. He holds that allowing artists to dictate the graphics creates the possibility of loss understanding of the "real news actually in the data" (87). Yet, he hesitates to set up his principles as any sort of verbal authority which would be in danger of dominating how we see. The eyes should be the final determinant of design success. What is sought is only "the clear portrayal of complexity" (191).

Bear in mind that Tufte discusses graphics to portray data or statistics themselves. The sort of infoviz schemes here in two and three dimensions portray information about the data—that is, how data are related to other data, or how much data are on a particular topic, and so on. His comment that "the number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data" (77) hints that he would not support the infoviz abstractions. To overlay his practicalities over the abstraction of infoviz interfaces exposes their incompatibility. He does mention a few principles that could pertain: Show data variation, not design variation. To draw an imprecise conclusion, one might say he suggests that interfaces do not need to be innovative. In essence, Tufte is concerned with visualization based on the data, and fi-

nally, on the eye. Nonetheless, his principles are cited in the context of infoviz regularly, where perhaps they do not belong.

Shneiderman and Plaisant (2005) are concerned with information visualization, or the abstract representation information about the data. Their principles seem more useful in searching for an answer to the question of which infoviz scheme works best. They write (2005, 102): "the central problem for human computer interaction researchers is to develop theories and models. Their eight golden rules of interface design have been refined over two decades to include (75-76): strive for consistency, offer informative feedback, prevent errors, permit easy action reversal, and cater to universal usability. Burkhard (2005, 252) acknowledges their guidelines and, based on his experience in a knowledge visualization company, adds some of his own to include: avoid decoration and avoid distracting your audience and prevent misinterpretation.

Infoviz theory and infoviz usability studies are related in that theory might profitably be grounded in empirical evidence derived from usability studies. Fundamental to usability studies are how we should evaluate, and what questions we should ask of users. Basic evaluations are often two-part: which systems are preferred, and given similar tasks, which "perform" better. Setting up a usability experiment is beyond the scope of this paper, but such an experiment, or even an outline of the experimental method to be followed, would be a noteworthy addition to the field.

7. Evalution for improvement

Experiments of single infoviz interfaces and experiments comparing different types of infoviz interfaces are not performed with any regularity. This section culls experimental findings on aspects of information visualization for digital collections in the areas of making categories, and in choice of visualization to show those categories. There is not yet enough evidence for an infoviz theory that would support the use of one scheme over another, but some of these findings are considered in more detail in the following section. In short, this discussion remains preliminary in the hope that other researchers will continue the inquiry.

7.1 Categorization

Many papers regarding information visualization in digital libraries assume that clustering has been successful. Some reported that classification had been done "accurately" (Goren-Bar and Kuflik 2005, 345) and (Krowne and Halbert 2005, 249); another that classification had been done "successfully" (Chen, Houston, Sewell and Schatz, 1998, 582). What appears to be meant by "successful" is that the documents were grouped by machine as a person might have grouped them. That is, there were no gross errors, such as documents about chicken clucking falling into the category of space exploration. Krowne and Halbert (2005) explain that classification accuracy is a well-known kind of metric. In the opinion of Geffner, Agrawal, El Abbadi and Smith (1999, 199): "Classification hierarchies provide an organization to data that is uniquely suited to digital library browsing because they organize collections from a user's point of view." However, a monolithic "user point of view" does not exist, so such clustering cannot be successful objectively.

People choose categories differently. The work of Furnas, Landauer, Gomez and Dumais (1987, 966) showed that the likelihood that many will assign the same name to the object is less than one in five. Between cultures, there might not be direct translations for words as basic as the colors seen by the human eye, or family relationships (Olson 1999). Within a culture, a word might alter its meaning over time (Lemesianou 2003). Further, different individuals within a culture might, for the same category, choose different words. Tuominen, Talja and Savolainen (2003, 563) seem to have it right: "Classification languages are supposed to reflect an order actually existing in the essence of things. However, this is not the case. Rather, they neutralize and conceal the inherent messiness of reality." Classifiers or automatic systems must group objects into the categories available. The categories are created based on what is in a collection at a given time, and subject to exceptions when new objects are added to a collection.

One study of user search behavior in a digital library found that users preferred scanning results of their keyword search to scanning the pre-made clusters (Das Neves and Fox 2000, 109). Nonetheless, the authors speculate that the users' preference might have been the result of the given task, and that a different search task might have resulted in users preferring to browse clusters. Some argue that classification systems limit the open, creative process of

finding information. But at least in the online environment, finding information freely is usually an option for a search by keyword. And large amounts of information contained in a digital library might well be too large in practical terms to be broken into smaller logical categories manually. For practical purposes of knowledge organization, Adams has a point in maintaining that automatic clustering enhances creativity by forming associations that would not otherwise be apparent (Adams 2001, 11). No classification system can be objectively accurate because the categories are not inherent unto themselves, but are created by people. Lack of ultimate objectivity does not imply that category-making should be abandoned, however. That would amount to rejecting a potentially useful tool because of fine-grained deficiencies. Categories are adequate as long as users understand what is meant by them.

7.2 Visualization

Preliminary findings on visualizations have been discouraging. Evaluation of infoviz interfaces generally is based on analysis of some specific task or tasks. For example, the interface of one three-dimensional visualization environment for a digital library initially "did not appear straightforward to all users" (Silva, Sánchez, Proal and Rebollar 2003, 156). A study of graphical overviews found that users thought the clustering confusing and the graphics less clear than text (Hearst 1999, 274). A study of SOMs found that users tend to get lost when browsing multi-level SOMs and prefer a conventional text-based alphabetic hierarchy (Chen, Houston, Sewell and Schatz 1998). Another recent study also found that most users preferred word lists over visualizations (Cole et al. 2005). And a casual survey of Rutgers University doctoral students in a fall 2005 information science seminar showed that students preferred text-based lists to 2D and 3D visualizations.

Morse and Lewis (2000) de-featured visualization retrieval displays from different systems so that they could compare not existing systems but rather the usability of the retrieval visualization on which the systems depend. They used 195 participants and a newswire document set from the TREC collection. (TREC=Text REtrieval Conference, an international conference started in the early 90s that provided large text collections on which participants conducted experiments in information retrieval.) The Morse and Lewis experiments compared 5 displays of information retrieved. These they called word (a list), table, icon (a display which looks like a simple bar chart), graph (with x and y axes) and spring (a simplification of a system called VIBE with tiles in bars that denote the weight of document with respect to the query term). The search question tasks varied in difficulty, which affected participants' error rates and execution times. The Morse and Lewis findings were that error rate was about the same independent of display used, but that participants worked somewhat faster using icon and spring displays than they did with the other types. When asked about preference, participants liked icons the most.

Fast and Sedig (2005) consider developing representations that are contextually appropriate to be a key challenge for digital libraries. Whether interfaces related to digital collections would be more easily understood by users remains to be tested. It is unclear what the optimum visualization should be specifically, although a meta-analysis of empirical studies of infoviz interfaces showed that users perform better with visual-spatial interfaces than with traditional interfaces. This meta-analysis showed that users do better when visualizations are simpler (Chen 2004 [1999], 179).

8. Toward a Theory of Information Visualization

It is shown in the previous section that some users have difficulty understanding visualization. Is that difficulty a factor of what people see (content)? How people see (color and the eye)? What people think? Or how they think? Each of these four factors is considered briefly below. If it could be determined *why* it is difficult for users to comprehend visualizations, we might be in a better position to design with clarity.

8.1 What people see

Many believe incorrectly that more concepts can be communicated visually than via other channels. This has been called the "Illusion of Visual Bandwidth" (Varakin, Levin and Fidler, 2004, 412). Perhaps it is this illusion that underlies many developers' overestimations of user ability to appreciate visual awareness (411). The authors conclude that "…visual objects do not necessarily have an advantage over text in terms of allowing a user to maintain awareness of multiple visual cues" (415). Based on this, one could

suggest that, to be successful, visualizations should lack intricate detail that could be easily overlooked. That is, visualizations should be simple. This was supported by evidence from the meta-analysis cited above (Chen 2004, 179).

8.2 How people see

Some researchers have tested whether it is the extra effort drawing the eye across the visualization that slows performance on such systems (Murata and Furukawa 2005). Two experiments were performed in which display characteristics were altered to require different eye movement. But experimental findings did not support the model. In quest for an alternate model, Murata and Furukawa speculate that memory affects performance. It might be that each visualization is an unknown and must be newly processed, thus slowing reaction time. This is similar to the suggestion of Dillon (2000), that people make sense of information displays by applying existing knowledge to new information. From this, one could suggest that, to be successful, visualizations should appear familiar to the user. In other words, visualizations should be reminiscent of an object seen or interface used before.

8.3 What people thin?

Dillon recommends that the optimum display be compatible with human tendencies of thought. This idea is echoed in what Chabris and Kosslyn (2005, 54) call the Representational Correspondence Principle, or what Tversky et al. (2007, 56) call the Principle of Congruence. The principle states that diagrams should show information in a form corresponding to the mental representation (Chabris and Kosslyn 2005, 54). Börner, Chen and Boyack reach the same conclusion (2003, 238).

The problem of what humans see in the mind is not answered simply. Rudolf Arnheim (1969, 100) believed that deep thought is imageless. Pylyshin (2003, 428) also believes that we do not know the format of thought. On the format of thought, he maintains that thoughts of intellects including Einstein, Maxwell, Faraday and others were non verbal. This coincides with evidence provided by Gardner (1983, 102 and 148) that the deepest thought is not experienced in that discipline's medium. While "in*sight*" suggests seeking, it is only metaphorical (Arnheim 1969, 101). Instead, thought images come at moments of subconscious thought in dreams. The images we do "see" in the mind seem to be partial, or selective in character, and memories of that which we have seen before (Arnheim 1969, 103 and 105). Would this suggest that visualizations are bound to be less effective than text, with which we are more familiar?

8.4 How people thin?

Explanations for why some have difficulty with visualizations focus on factors such as spatial ability and memory. Chen conducted a meta-analysis on empirical studies of visualizations in an attempt to find patterns. He located 35 such studies published between 1991 and 2000, but for the purposes of the meta-analysis, could use results from only 6 of the 35. He compared the extent to which users with stronger memory perform better in information retrieval using visualizations, and the extent to which users perform better with or without visualizations. His results analyzing 6 studies (2004, 178-179) showed that those with stronger memory do better in visualization retrieval, and results from 5 studies showed that user perform better with visual-spatial information retrieval interfaces than with traditional retrieval interfaces.

8.5 Towards a Theory of Information Visualization: Predictability

These explanations of what people see, how they see, what they think and how they think lend credence to some of the principles for designing infoviz interfaces such as 'strive for consistency' and 'strive for simplicity'. Even more specific guidelines along with comparative usability tests of specific designs might encourage designers to repeat what is prescribed and what is proved well-liked by users. In the world of gadgets, predictability trumps innovation because the more familiar the look and interaction, the greater will be the system usability.

9. Implications for Further Research and Development

9.1 Creating categories

Different people often choose different words for the same category (Furnas et al., 1987). Suppose a

system could allow users to add their own categories, which the system then could use to group documents on-the-fly. Documents would not be pregrouped, but would form groups as needed. A goal of large digital library and database projects is often interoperability with other digital libraries and databases. How can diverse knowledge domains co-exist if clustering is based on inter-document similarities? For example, some documents will cluster chronologically, others will cluster geographically, others according to related subjects. We will want to combine such clusters, but how to make sense of this semantically and visually awaits research.

9.2 Showing categories

Users might apprehend graphics better if the designs were clearer, although what "clear" means in this context is unclear. More obvious would be a visualization that permitted customization for ergonomics or attractiveness. Font size and color, for example, could be attributes that a user alters according to preference, while 3D schemes might translate into tactile representations for the visually impaired. Determining which features enhance user experience, and whether the visualization format should align with knowledge domain are possible avenues for future research. According to Wan (2006) the current trend in visualization is in 3D space and virtual reality. Wan reports that the 3D structures are more difficult to implement than the 2D, so the 2D structures developed in the 1990s continue to be used widely (Wan 2006, 93).

An alternative to developing new graphical schemes might be to adopt commercial software with information visualization properties. Two software products that have met with market success in their own right are *Grokker* and *Inxight*. *Grokker*, by Groxis, enlivens its web search interface with shapes, colors and animation that resonates with a school audience. *Inxight* powers the *Collections at a Glance* section of the National Science Digital Library (http://nsdl.org/browse/ataglance/browseBySubject.html).

9.3 Interaction design

Some of the schemes above, such as the hierarchical list and star map, are intended for browse interfaces. Other schemes such as the tree map, SOM and domain landscape are intended for display of results. Placement of a visualization scheme could enhance its utility. Infoviz overviews should not be tucked away among various menu choices but rather should be situated on an opening screen, perhaps with a keyword box as a search alternative. Think about a physical library, where visitors can estimate the relative size of a collection by the number of shelves or floors it spans.

Shen, Vemuri, Fan, Torres and Fox (2006) created a prototype digital library visualization in which, under certain circumstances, browse can be converted to search, and vice versa. In the evaluation phase, some users appreciated the ability to save a navigation path, and to search within a browse context. The intermingling of search and browse functions is another possible avenue of research. Baudel remarks (2006, 67) that interaction in existing infoviz packages is limited to navigation features. He would add editing features to edit groups, clone or remove objects, or add or remove attributes. Considering how a digital collection will be used will allow developers to outfit it with a more useful feature set.

9.4 Predictability

Again, it is wise to recall that predictability engenders usability. The designs that proved best—presumably designs that are simple or familiar—should be used again and again, so that users will become even more familiar and will be able to interpret them more reliably.

10. Conclusion

A range of information visualization schemes can be used to classify items in databases or digital libraries. We presently lack a theory of information visualization to suggest which scheme would be best. In approaching a theory, we have considered evaluations of infoviz interfaces and possible explanations for the lukewarm reception of the abstractions. Research and development of infoviz interfaces might best proceed by looking beyond the literature to concentrate on functioning digital libraries in an attempt to understand what works best and why.

References

Adams, Katherine C. 2001. Information architecture as narrative: the function and benefits of taxonomies. In *Internet Librarian 2001 Collected Presentation, Pasadena, California November 6-8,* 6-11. Medford, NJ: Information Today, Inc.

Amar, Robert and Stasko, John. 2004. A knowledge task-based framework for design and evaluation of information visualizations. In *IEEE Symposium on Information Visualization Octoaber 10-12, Austin, Texas*. Washington: IEEE Press, pp. 143-49.

Arnheim, Rudolf. 1969. *Visual thinking*. Berkeley, California: University of California Press.

Baudel, Thomas. 2006. From information visualization to direction manipulation: Extending a generic visualization framework for the interactive editing of large datasets. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology UIST '06, October 15-18, Montreux, Switzerland*. New York: ACM Press, pp. 67–76.

Belkin, N. J., Oddy R. N., and Brooks, H. 1982. ASK for information retrieval Part I. *Journal of documentation 38:* 61–72.

Börner, Katy, Chen, Chaomei, and Boyack, Kevin W. 2003. Visualizing knowledge domains. In Cronin, Blaise, ed., *Annual review of information science and technology* 37: 179–255.

Boyack, Kevin, Wylie, Brian N., and Davidson, George S. 2002a. Information visualization, human-computer-interaction, and cognitive psychology: domain visualizations. In Börner, Katy, and Chen, Chaomei, eds., *Visual interfaces to digital libraries*. Heidelberg: Springer, pp. 145-58

Boyack, Kevin, Wylie, Brian N., and Davidson, George S. 2002b. Domain visualization using *VxInsight* for science and technology management. *Journal of the American Society for Information Science and Technology* 53: 764–74.

Burkhard, Remo Aslak. 2005. Towards a framework and a model for knowledge visualization: synergies between information and knowledge visualization. In Tergan, S.-O. and Keller, T., eds., *Knowledge and Information Visualization, Lecture Notes in Computer Science 3426.* Heidelberg: Springer, pp. 238–55.

Chen, Chaomei. 1999. Visualizing semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management* 35: 401–20.

Chen, Chaomei. 2000. Domain visualization for digital libraries. In *Proceedings of the IEEE International Conference on Information Visualization, July 19-21, London, United Kingdom*. Washington: IEEE Press, pp: 261–67.

Chen, Chaomei. 2004. *Information visualization: beyond the horizon*. 2nd ed. London: Springer.

Chen, Chaomei, and Börner, Katy. 2002. Top ten problems in visual interfaces to digital libraries. In Chen, Chaomei, and Börner, Katy eds., *Visual interfaces to digital libraries.* Berlin: Springer, pp. 227–32.

Chen, Hsinchun, Houston, Andrea L., Sewell, Robin R. and Schatz, Bruce R. 1998. Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science* 49: 582–603.

Cole, Charles, Leide, John E., Large, Andrew, Beheshti, Jamshid, and Brooks, Martin. 2005. Putting it together online: information need identification for the domain novice user. *Journal of the American Society for Information Science and Technology* 56: 684–94.

Chen, Hsinchun, Schuffels, Chris and Orwig, Rich. 1996. Internet categorization and search: A machine learning approach. *Journal of visual communications and image representation* 7: 88-102.

Das Neves, Fernando A. and Fox, Edward A. 2000. A study of user behavior in an immersive virtual environment for digital libraries. In *Proceedings of the fifth ACM International Conference on Digital Libraries, June 2-7, San Antonio, Texas*. Washington: ACM Press, pp. 103–11.

Deng, Da, Zhang, Jianhua, and Purvis, Martin. 2004. Visualisation and comparison of image collections based on self-organised maps. In *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation*. 32: 97–102.

Dillon, Andrew. 2000. Spatial semantics: how users derive shape from information space. *Journal of the American Society for Information Science* 51: 521–28.

Fabrikant, Sara. 2000. Spatialized browsing in large data archives. *Transactions in GIS* 4: 65–78.

Fast, Karl V. and Sedig, Kamran. 2005. The INVENT framework: examining the role of information visualization in the reconceptualization of digital libraries. *Journal of digital information* 6 issue 3 article no. 362, 2005-08-08. http://jodi.tamu.edu/Articles/v06/i03/Fast/

Feng, Ling, Jeusfeld, Manfred A., and Hoppenbrouwers, Jeroen. 2005. Beyond information searching and browsing: acquiring knowledge from digital libraries. *Information processing & management* 41: 97–120.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. 1987. The vocabulary problem in human-system communication. *Communications of the ACM 30*(11): 964–71.

Gardner, Howard. 1983. *Frames of mind: the theory of multiple intelligences*. New York: Basic Books.

Geffner, Steven, Agrawal, Divyakant, El Abbadi, Amr, and Smith, Terence R. 1999. Browsing large digital library collections using classification hierarchies. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*. New York: ACM Press, pp. 195–201.

Goren-Bar, Dina, and Kuflik, Tsvi. 2005. Supporting user-subjective categorization with self-organizing maps and learning vector quantization. *Journal of the American Society for Information Science and Technology* 56: 345–55.

Greene, Stephan, Marchionini, Gary, Plaisant, Catherine, and Shneiderman, Ben. 2000. Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *Journal of the American Society for Information Science* 51: 380–93.

Hearst, Marti A. 1999. User interfaces and visualization. In Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier, eds., *Modern information retrieval*. New York: ACM Press, pp. 257–323.

Heer, Jeffrey, Card, Stuart K., and Landay, James A. 2005. *Prefuse*: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 2-7, Portland Oregon*, pp. 421–30.

Judelman, Greg. 2004. Aesthetics and inspiration for visualization design: bridging the gap between art and science. In *Eighth International Conference on Information Visualization, 14-16 July, London, England*. London: International School of New Media, pp. 245–50.

Kohonen, Teuvo. 1990. The self-organizing map. *Proceedings of the IEEE 7:* 1464–80.

Krowne, Aaron and Halbert, Martin. 2005. An initial evaluation of automated organization for digital library browsing. In *Proceedings for the Joint Conference on Digital Libraries*, *June 7-11, Denver, Colorado.* New York: ACM Press, pp. 246–55.

Lemesianou, Christine A. 2003. Sign matters: the shift in semantic landscapes of the sign 'generation x' through time. In Mokros, Hartmut B. ed., *Identity matters: communication-based explorations and explanations.* Cresskill, N.J.: Hampton Press, pp. 31-54.

Lesk, Michael. 2005 *Understanding digital libraries*. 2nd ed. San Francisco: Morgan Kaufmann.

Marshall, Byron and Madhusudan, Therani. 2004. Element matching in concept maps. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL '04) June 7-11, 2004, Tuscon, Arizona*. Washington, D.C.: IEEE Computer Society, pp. 186–7.

Marshall, Byron, Zhang, Yiwen, Chen, Hsinchun, Lally, Ann, Shen, Rao, Fox, Edward and Cassell, Lillian. 2003. Convergence of knowledge management and e-learning: The GetSmart experience. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, May 27-31, Houston, Texas*. Washington, D.C.: IEEE Computer Society, pp. 135–46.

Merkl, Dieter and Rauber, Andreas. 2000. Digital libraries—classification and visualization techniques. In *International Conference on Research and Practice, November 13-16, Kyoto, Japan,* pp. 434–38.

Morse, E. and M. Lewis. 2000. Evaluating visualizations: Using a taxonomic guide. *International journal of human-computer studies* 53: 637–62.

Murata, Atsuo and Furukawa, Nobuyasu. 2005. Relationships among display features, eye movement characteristics, and reaction time in visual search. *Human factors* 47: 598–612.

Mylopoulus, John. 1981. An overview of knowledge representation. In Brodie, Michael L. and Zilles, Stephen E. eds., *Proceedings of the 1980 Workshop on Data Abstraction, Databases and Conceptual Modeling. June 23-26, Pingree Park, Colorado*. New York: ACM Press, pp. 5-12.

Novak, Joseph D. 1998. *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Mahwah, N. J.: Lawrence Erlbaum Associates.

Novak, Joseph D. and Cañas, Alberto J. 2006. *The theory underlying concept maps and how to construct them*. Technical report IHMC CmapTools 2006-01, Florida Institute for Human and Machine Cognition. http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf.

Olson, Hope A. 1999. Cultural discourses of classification: Indigenous alternatives to the tradition of Aristotle, Durkheim, and Foucault. In

Albrechtsen, Hanne and Mai, Jens-Erik eds., *Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, November 1-5, Washington, DC*. Advances in classification research 10. Medford, N.J.: Information Today, pp. 91–106.

Phillips, Matthew, Rajkumar, Johnny, and Shao, Zhiyan. 2005. *CS Fall 2005 Project Report: Analysis of GetSmart Concept Maps.* http://pubs.dlib.vt.edu:9090/40/01/cs6604report_fall2005_cmapanalysis_final.pdf.

Polanco, Xavier, Francois, Claire, and Lamirel, Jean-Charles. 2001. Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach. *Scientometrics* 51: 267–92.

Pylyshyn, Zenon W. 2003. *Seeing and visualizing*. Cambridge: MIT Press.

Ramsey, Marshall C., Chen, Hsinchun., Zhu, Bin., and Schatz, Bruce R. 1999. A collection of visual thesauri for browsing large collections of geographic images. *Journal of the American Society for Information Science* 50: 826–34.

Roussinov, Dmitri G. and Chen, Hsinchun. 1998. A scalable self-organizing map algorithm for textual classification: a neural network approach to thesaurus generation. *Communication, cognition and artificial intelligence* 15, no. 1-2: 81-111. 1998. http://dlist.sir.arizona.edu/460/01/A_Scalable-98.htm.

Saracevic, Tefko. 2004. Evaluation of digital libraries: An overview. Presentation at the *DELOS WP7 Workshop on the Evaluation of Digital Libraries, October 4-5, Department of Information Engineering, University of Padua, Italy*. http://www.scils.rutgers.edu/~tefko/articles.htm.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34: 1–47.

Shen, Rao, Vemuri, Naga Srinivas, Fan, Weiguo, Torres, Ricardo da S. and Fox, Edward A. 2006. Exploring digital libraries: Integrating browsing, searching, and visualization. In *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries, June 11-15, Chapel Hill, North Carolina*. New York: ACM Press, pp. 1-10.

Shneiderman, Ben. 2006. *Treemaps for space-constrained visualization of hierarchies*. http://www.cs.umd.edu/hcil/treemap-history/.

Shneiderman, Ben, Feldman, David, Rose, Anne, and Grau, Xavier Ferré. 2000. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, *June 2-7, San Antonio, Texas*. New York: ACM Press, pp. 57–66.

Shneiderman, Ben and Plaisant, Catherine. 2005. *Designing the user interface: strategies for effective human-computer interaction*. 4th ed. Boston, MA: Pearson.

Silva, Nabani N., Sánchez, J. Alfredo, Proal, Carlos and Rebollar, Christian. 2003. Visual exploration of large collections in digital libraries. In *Proceedings of the Latin American Conference on Human-computer interaction, August 17-20, Rio de Janeiro, Brazil*. ACM International Conference proceeding series 46. New York: ACM Press, pp. 147–57.

Sumner, Tamara, Ahmand, Faisal, Bhushan, Sonal, Gu, QIanyi, Molina, Francis, Willard, Stedman, Wright, Michael, Davis, Lynne, and Janée, Greg. 2005. Linking learning goals and educational resources through interactive concept map visualizations. *International journal of digital libraries* 5: 18–24.

Tufte, Edward R. 2001. *The visual display of quantitative information*. 2nd Ed. Cheshire, Connecticut: Graphics Press.

Tuominen, Kimmo, Talja, Sanna, and Savolainen, Reijo. 2003. Multiperspective digital libraries: the implications of constructionism for the development of digital libraries. *Journal of the American Society for Information Science and Technology 54:* 561–69.

Tversky, Barbara, Agrawala, Maneesh, Heiser, Julie, Lee, Paul, Hanrahan, Pat, Phan, Doantam, Stolte, Chris, and Daniel, Marie-Paule. 2007. Cognitive design principles for automated generation of visualizations. In Allen, Gary L. ed., *Applied spatial cognition: from research to cognitive technology*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 53-73.

Varakin, D. Alexander, Daniel T. Levin, and Fidler, Roger. 2004. Unseen and unaware: implications of recent research on failures of visual awareness for human-computer interface design. *Human Computer Interaction* 19: 389–422.

Wan, Gang. 2006. Visualizations for digital libraries. *Information technology and libraries* 25: 88–94.