SANDIA REPORT
SAND2000-3100
Unlimited Release
Printed December 2000

Visualization of Information Spaces with VxInsight

B. N. Wylie, K. W. Boyack, G. S. Davidson, and D. K. Johnson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

SAND2000-3100
Unlimited Release
Printed December 2000

Visualization of Information Spaces with VxInsight

B. N. Wylie
Data Analysis & Visualization Department

K. W. Boyack, G. S. Davidson, D. K. Johnson
Evolutionary Computing/Agents Department

Sandia National Laboratories
P.O. Box 5800 MS-0318
Albuquerque, NM 87185-0318

Abstract

The information age allows people to find, extract, and gather information in astronomical proportions. Commercial databases, the World Wide Web, and data warehouses offer the promise of knowledge and insight if one can harness the overwhelming volume of raw information. Currently, with traditional tools, the task of extracting knowledge from large databases is exceedingly difficult. A number of data mining tools exist to answer fairly specific questions about data collections, but more often than not, an analyst doesn't even know what questions to ask. Also, without the understanding of the global 'picture', specific questions and observations cannot be placed within the proper context. Unfortunately, existing tools are not well suited for studying large-scale structure or interacting with data in an exploratory manner.

One of the key shortcomings of information management systems is in the nature of the user interface. Interpreting textual information is a sequential and slow process, which imposes a limit on the speed with which information can be assimilated and processed. Text is excellent for conveying detailed information, but it is poorly suited for conveying relationships or trends or for getting an overview of a set of data. For these tasks, graphical displays are more effective and have become essential elements of scientific research (Tufte, 1992). The visual system is incredibly effective at identifying trends and patterns, and detecting anomalies in large sets of data. These innate skills have the potential to significantly alleviate the problem of information overload. For this potential to become a reality, graphical interfaces must be developed for all types of abstract information. These interfaces must be carefully designed to leverage the inherent talents of our visual systems. However, they must also be modest, only exploiting visualization where appropriate, because many kinds of information are best conveyed through other channels. Our attempt at these goals is a tool that provides a simple visualization environment in which the user can intuitively exploit the power of relational databases.

VxInsight™ is a powerful and flexible tool for exploring data collections (Davidson et al. 1998). It works by transforming the data into an intuitive visual format that is easy to interpret and allows natural navigation. Millions of years of evolution have equipped us with extraordinary powers to visually spot trends and patterns, to identify outliers and to detect relationships. VxInsight exploits this capability by presenting the data as a landscape, a familiar representation that we are adept at interpreting and which allows very large data sets to be represented. The landscape representation conveys significant information about the implicit structure of the data, providing context for the analyst's exploration and queries of the database.

Contents

I. Introduction 3
II. Related Work 3
III. VxInsight Tool 4
IV. VxInsight ordination routine (VxOrd) 6
V. User environment 8
VI. Analysis methods 9
VII. Applications 10
VIII. Summary and Future Work 15
IX. Figures 17
X. References 23
XI. Appendix 25

I. Introduction

The information age allows people to find, extract, and gather information in astronomical proportions. Commercial databases, the World Wide Web, and data warehouses offer the promise of knowledge and insight if one can harness the overwhelming volume of raw information. Currently, with traditional tools, the task of extracting knowledge from large databases is exceedingly difficult. A number of data mining tools exist to answer fairly specific questions about data collections, but more often than not, an analyst doesn't even know what questions to ask. Also, without the understanding of the global 'picture', specific questions and observations cannot be placed within the proper context. Unfortunately, existing tools are not well suited for studying large-scale structure or interacting with data in an exploratory manner.

One of the key shortcomings of information management systems is in the nature of the user interface. Interpreting textual information is a sequential and slow process, which imposes a limit on the speed with which information can be assimilated and processed. Text is excellent for conveying detailed information, but it is poorly suited for conveying relationships or trends or for getting an overview of a set of data. For these tasks, graphical displays are more effective and have become essential elements of scientific research (Tufte, 1992). The visual system is incredibly effective at identifying trends and patterns, and detecting anomalies in large sets of data. These innate skills have the potential to significantly alleviate the problem of information overload. For this potential to become a reality, graphical interfaces must be developed for all types of abstract information. These interfaces must be carefully designed to leverage the inherent talents of our visual systems. However, they must also be modest, only exploiting visualization where appropriate, because many kinds of information are best conveyed through other channels. Our attempt at these goals is a tool that provides a simple visualization environment in which the user can intuitively exploit the power of relational databases.

VxInsight™ is a powerful and flexible tool for exploring data collections (Davidson et al. 1998). It works by transforming the data into an intuitive visual format that is easy to interpret and allows natural navigation. Millions of years of evolution have equipped us with extraordinary powers to visually spot trends and patterns, to identify outliers and to detect relationships. VxInsight exploits this capability by presenting the data as a landscape, a familiar representation that we are adept at interpreting and which allows very large data sets to be represented. The landscape representation conveys significant information about the implicit structure of the data, providing context for the analyst's exploration and queries of the database.

II. Related Work

Various efforts to map the structure of science have been undertaken over the years. Science mapping studies are typically focused at either the macro or micro level. At a macro-level such studies (Bassecoulard & Zitt, 1999; Nederhof & van Wijk, 1997) seek to determine the basic structural units of science and their interrelationships. Some macro level studies also allow exploration of fine-scale structure underlying the global networks (Small, 1999). The majority of studies are performed at the discipline or specialty level (Leydesdorff, 1994; Noyons & van Raan, 1998; McCain, 1998; Spasser, 1997), and seek to inform science policy and technical decision makers. Studies at both levels probe the dynamic nature of science, and the implications of the changes. Alternate approaches with more applied goals include textual data mining (Losiewicz, 2000) and database tomography methods (Kostoff, 1999), and are usually applied at the discipline level.

A variety of databases and methods have been used for these studies. Primary among databases are the Science Citation Indices (SCI and SSCI) from the Institute for Scientific Information (ISI), which have gained widespread acceptance for bibliometric studies. Maps are most often based on computed similarities between journal articles (McCain, 1998; Small, 1999), co-occurrence or co-classification using keywords, topics, or classification schemes (Nederhof & van Wijk, 1997; Noyons & van Raan, 1998; Spasser, 1997), or journal citation patterns (Bassecoulard & Zitt, 1999; Leydesdorff, 1994).

Once a similarity matrix is defined, algorithms are used to cluster the data. Common clustering methods for producing maps include multidimensional scaling, heirarchical clustering, k-means algorithms, and self-organizing maps. The standard mapping output is a circle plot where each cluster is represented by an appropriately sized circle. Links between circles provide relationship information. Traditionally, map outputs have been paper-based and only resolve structure at a few discrete levels. However, in recent years, several systems have been reported that use a computer display and allow some navigation of the map space.

SENTINEL (Fox et al., 1999) is a Harris Corporation package that combines a retrieval engine using n-grams and context vectors for effective query with a visualization system called VisualEyes™. The visualization tool allows the user to interact with document clusters in a three-dimensional space. Chen (1999) uses a VRML 2.0 viewer in conjunction with Generalized Similarity Analysis to display papers (as spheres) and the Pathfinder linkage network connecting them which has been calculated from a co-citation analysis.

Self-organizing maps have been used in many venues, including the organization of document spaces (Honkela et al, 1998). These maps are used to position documents, and then display them in a two-dimensional contour-map-like display in which color represents density. Peak labels can be generated automatically, and some limited navigational and retrieval capabilities are provided.

Two packages that are more similar to Sandia's VxInsight are SCI-Map developed by ISI (Small, 1999), and the SPIRE suite of tools that originated at Pacific Northwest National Laboratory (Hetzler et al., 1998; Wise, 1999). SCI-Map uses a hierarchically nested set of maps to display the document space at varying levels of detail. This nesting of maps allows drilling down to subsequent levels. Each map is similar to the traditional circle plot, where the size of the circle can indicate the density of documents contained in the circle, or some measure of importance. Relationships at each discrete level are indicated by links between circles.

Like VxInsight, SPIRE maps objects to a two-dimensional plane so that related objects are near each other, and provides tools to interact with the data. SPIRE has two visualization approaches. In the Galaxies view, documents are displayed as a scatter plot. This interface allows drilling down to smaller sections of the scatter plot, and provides some summarization tools. In the Themescape view, a high-level terrain display, similar to that in VxInsight, is used. Themescape visualizes specific themes as mountains and valleys, where the height of a mountain represents the strength of the theme in the document set. However, unlike VxInsight, this view is static, and allows no zooming in or out of the terrain to get more detailed information. Thus, SPIRE does not provide the continuous, multi-resolution viewing that is essential for revealing the inherent structure of an information space at multiple levels of detail.

III. VxInsight Tool

VxInsight overlays the 2-D plane with a 3-D virtual landscape that looks like a mountain range. The height of a mountain is proportional to the density of objects beneath it. This 3-D environment is readily explored because there is only a small cognitive step between seeing the presented images and then exercising our innate human expertise in navigating through real landscapes. As can be seen in Figure 1, the interface is straightforward and intuitive. Ease of data exploration remains one of the guiding principles of the VxInsight design.

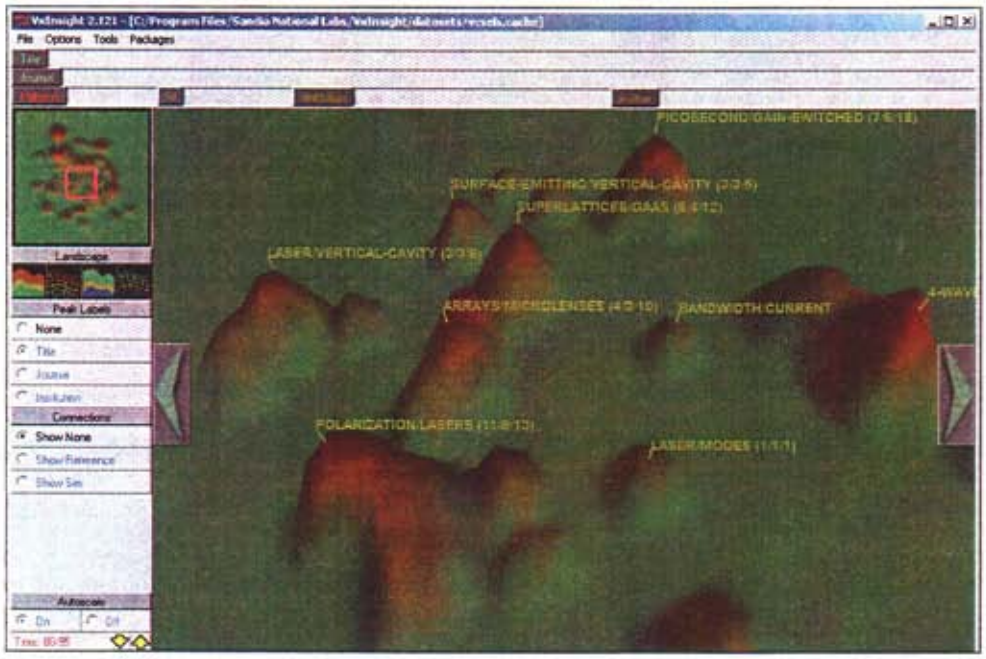


Figure 1: VxInsight User Interface

VxInsight allows one to see the big picture, to zoom in and see the details right down to the network of relationships among the data elements. If the database represents data that has accumulated through time, VxInsight also allows the analyst to see the terrain representations evolve temporally as the data set increases through its collection lifetime. In our experience, these views of the data can be highly informative. But the real power of VxInsight becomes apparent when the analyst begins to ask questions using the Standard Query Language (SQL) interface to the database. Typically, the analyst will compose a Boolean query that is

passed back to the database engine. VxInsight uses the response from the database engine to visually mark the data elements that match the query. It does so by coloring spots (on the 3-D terrain representation) above the matched data elements. The results of each query are shown with different colors. This places query results in the context of the overall information terrain, and in relation to other queries, thereby providing a powerful way to answer questions. This representation shows not only the matching contents of the database, but also says important things about the intersection between the analyst's question and the implicit structure of the database. Hypotheses can be formulated, tested and revised, and the answers can suggest new questions -- all in real time. It is this participatory interaction with the data that gives VxInsight its power.

VxInsight provides a visual mechanism for browsing, exploring and retrieving information from a database. The graphical display conveys information about the relationship between objects in several ways and on multiple scales. In this way, individual objects are always observed within a larger context.

For example, consider a database consisting of a set of scientific papers. Imagine that the papers have been organized in a two dimensional geometry so that related papers are located close to each other. Now construct a landscape where the altitude reflects the local density of papers. Papers on physics will form a mountain range, and a different range will stand over the biological papers. In between will be research reports from biophysics and other bridging disciplines. Now, imagine exploring these mountains. If we zoom in closer, the physics mountains will resolve into a set of sub-disciplines. Eventually, by zooming in far enough, the individual papers become visible. By pointing and clicking you can learn more about papers of interest or retrieve their full text.

Although physical proximity conveys a great deal of information about the relationship between documents, you can also see which papers reference which others, by drawing lines between the citing and cited papers. For even more information, you can choose to highlight papers by a particular researcher or a particular institution, or show the accumulation of papers through time, watching some disciplines explode and other stagnate.

VxInsight is a general-purpose tool, which enables this kind of interaction with wide variety of relational data: documents, patents, web pages, and financial transactions are just a few examples. The tool allows users to interactively browse, explore and retrieve information from the database in an intuitive way.

Several basic principles guided the design of the tool.

1. The human visual system excels at identifying patterns, trends and anomalies.
2. A human can better interpret data with presentations built on familiar metaphors.
3. A useful tool allows easy and intuitive navigation that provides both detailed and high-level views.
4. Comprehension of large datasets is facilitated by the ability to explore the data at different resolutions, for instance, by zooming into and out of the representation. To enhance interpretability, the resolution should vary continuously.
5. Since no single representation is optimal for all types of questions, a tool should contain a tool kit of different display and interaction mechanisms.
6. Graphical displays should have compact representations to facilitate interactive frame rates and access over a network. However, simplicity must not prevent the tool from being applicable to large datasets.

In VxInsight, we use geometric proximity as a metaphor showing the relatedness of two objects, and present the result as 3-D landscape. This is a very intuitive mechanism, but it has significant limitations. Databases typically encode many different kinds of relationships, so proximity in a low-dimensional space is insufficient to fully capture all of the complexity in the relationships. Although placing the objects in a higher-dimensional geometric space would allow proximity to convey more information, users have great difficulty interpreting and navigating higher dimensional spaces.

IV. VxInsight ordination routine (VxOrd)

VxOrd is used in conjunction with the VxInsight application to generate the spatial location of data objects. The location of data objects is based on the similarities between the objects. In the figure below VxOrd has generated coordinates for a particular group of data. Objects with a high similarity are clustered together; objects with little or no similarity are separated on the map.

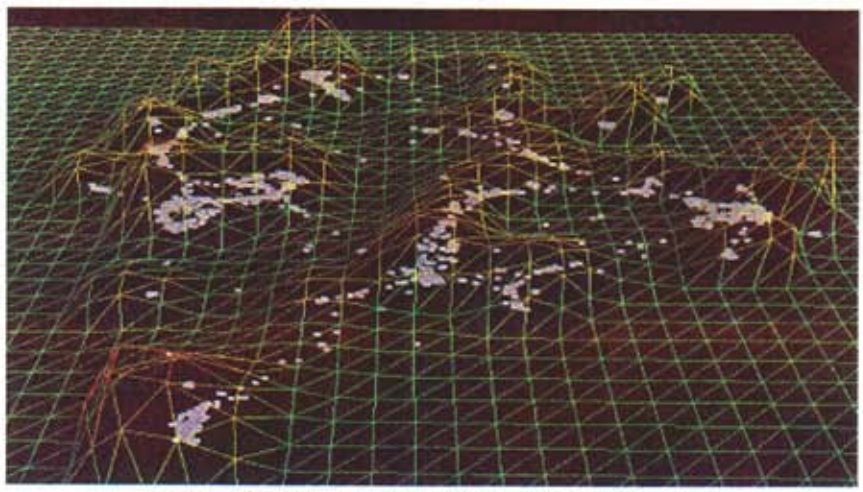


Figure 2: Clustering of data based on VxOrd calculation

At the highest level this algorithm tries to place similar objects close together and dissimilar objects far apart. The example shown in Figure 3 demonstrates this basic principle. Here, the process is achieved by moving the objects randomly around the solution space via a technique similar to 'simulated annealing'. The criteria for moving a node is the minimization of energy given by Equation 1:

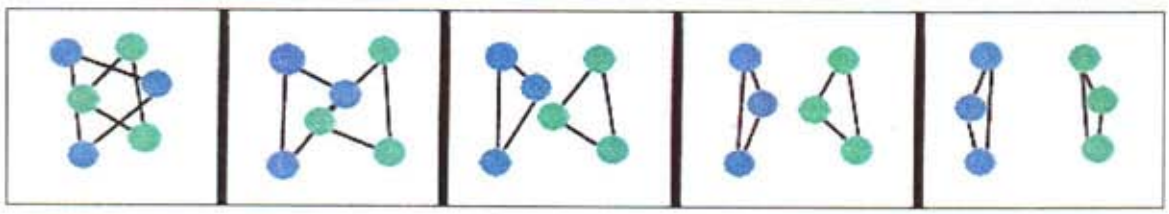


Figure 3: The process of moving similar items close together and simultaneously pushing dissimilar objects away.

$$E_{x,y} = \left[\sum_{i=0}^{n} \left(w_i \times l_i^2 \right) \right] + D_{x,y}$$ ***Equation 1***

$E_{x,y}$ = The energy of a node at a specific x, y location
w_i = The similarity between this node and the node connected by edge i.
l_i^2 = The Euclidean distance squared between this node and the node connected by edge i.
$D_{x,y}$ = The density measure with respect to the area around point x,y.

Simplified pseudo code is given to illustrate the process:

```
for(j=0; j<number_of_iterations;j++) {

      // Now randomly move nodes to see if we can minimize their energy
      for(I=0;I<number_of_nodes;I++) {

        N = node[I];

        // Move node
```

```
        N.moveNode(randomX, randomY);

        // New energy is less so take the move
        if (N.newEnergy < N.oldEnergy)
                N.takeMove();
        else
                N.moveBack();
    }
}
```

Implementation of density field to achieve order N running time

The function $D_{x,y}$ can be computed in many different ways. The brute force approach is to base $D_{x,y}$ on the distance to all other nodes. Since each node would have to check it's position against all other nodes this brute force approach would take N comparisons (where N = number of nodes in graph) for each determination of $D_{x,y}$. All nodes must compute $D_{x,y}$ when determining their energy at a specific location x,y, so in total the algorithm would require $O(N^2)$ running time.

For large datasets an $O(N^2)$ running time would be prohibitively computationally expensive. We have developed a specific method of computing $D_{x,y}$ which allows each node to determine an approximate value for this term in constant time O(1), thereby reducing the total running time to O(N).

The algorithm called the 'Density Field' algorithm is implemented by having each node place an energy footprint onto a 2 dimensional array (density field). The energy footprint may be any function in 2 space, currently the current implementation uses a circle with some radius *r*. The energy footprint is highest in the center of the circle and diminishes with the radius. An alternative would be a 2 dimensional gaussian function or a 2 dimensional sinusoidal or any other 2 dimensional function familiar to those skilled in the art. The density field is constructed by summing each node's energy footprint. Having each node contribute an energy footprint to the density field requires O(N) time.

Now that the density field is constructed, a node can determine an approximate value to the $D_{x,y}$ term by a constant time lookup into the density field at location x,y. Having each node look up the value of the density field at its current location requires O(N) time.

The utilization of the density field algorithm requires both the time for construction and the time for lookup.

Determination of Density Term = Construction [O(N)] + Lookup [O(N)]
Determination of Density Term = O(N) + O(N)
Determination of Density Term = 2 * O(N)

Determination of Density Term = O(N) (in asymptotic notation constants are ignored)

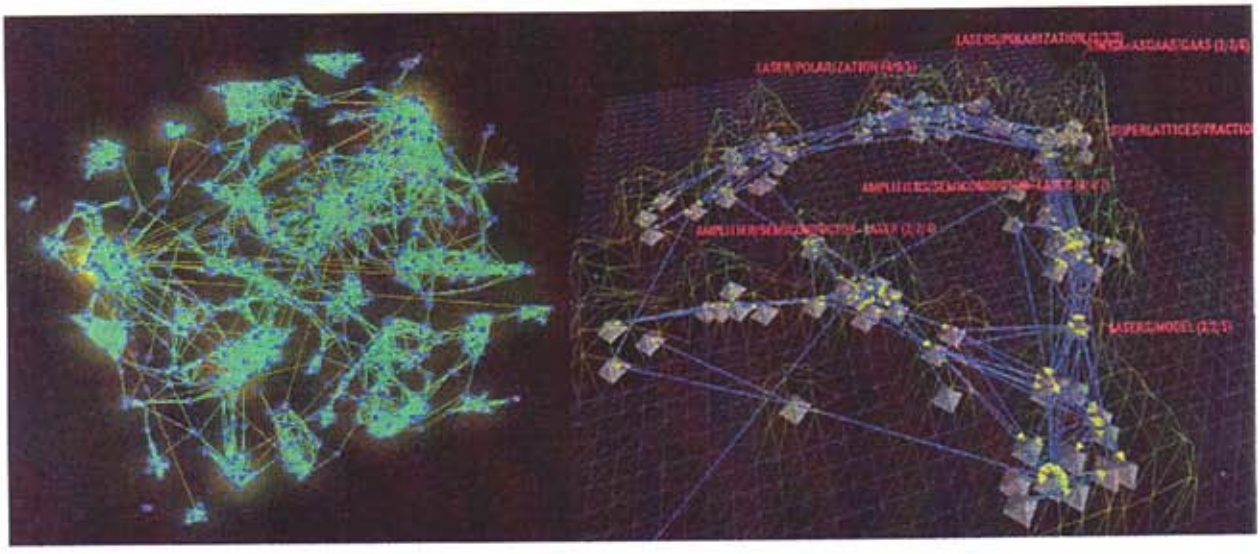


Figure 4: A view of a clustered graph with density field on the left, same ordination at different resolution on right.

V. User environment

The environment provided for the user is extremely important in the exploration of data. To rapidly interact with data on many conceptual levels requires an intuitive and unencumbered environment. VxInsight provides this environment by presenting the user with a large interaction window in which simple navigational constructs allow the user to traverse information space.

As with other controls, the zooming mechanism is simple and straightforward. The analyst picks an area of interest and clicks on it to zoom in. Holding the first button down provides a fluid transition to higher viewing resolutions. To zoom out, simply press the second mouse button. A typical analyst will change the viewing resolution quite often during a session of data exploration.

VxInsight provides four options for rendering mountain landscapes. The most popular is the transparent wire frame rendering, depicted in Figure 2, which lets the analyst see the density of data elements below the mountains and the connection networks on higher levels. As shown in Figure 1, the user can also view the terrain as solid mountains, which can displayed in two different color contrast schemes.

To explain the nature of the landscape, VxInsight dynamically generates labels for the most significant mountain peaks. It does this by looking for the most common words in a particular database field among the objects comprising a mountain. The analyst can dynamically change the field being used for labels. The peak label provides a rough estimate of the content of a mountain. The label displays the two most common words along with the number of times they occurred and the number of items sampled beneath the peak. The text processing required for peak labeling can be time consuming if a large number of objects are involved. To address this problem, VxInsight allows the user to specify a maximum number of objects to include in the processing. This results in only a sampling of objects being considered, trading off label quality for run-time. An example of the peak information strings can be seen in Figure 1.

VxInsight can also display directional links between data objects. A dataset on 'personnel interactions' may have dynamic connection fields such as 'Phone calls' and 'Lunches'. If the objects are scientific papers, the connections could be citation links. When the user selects one of the connections for display, directional links between objects will appear as can be seen on the right side of Figure 4. These links give the analyst information about the structure of the data. On the highest level there may be so many connections that not much information can be extracted, so this option is mainly used when the current resolution contains a manageable number of objects.

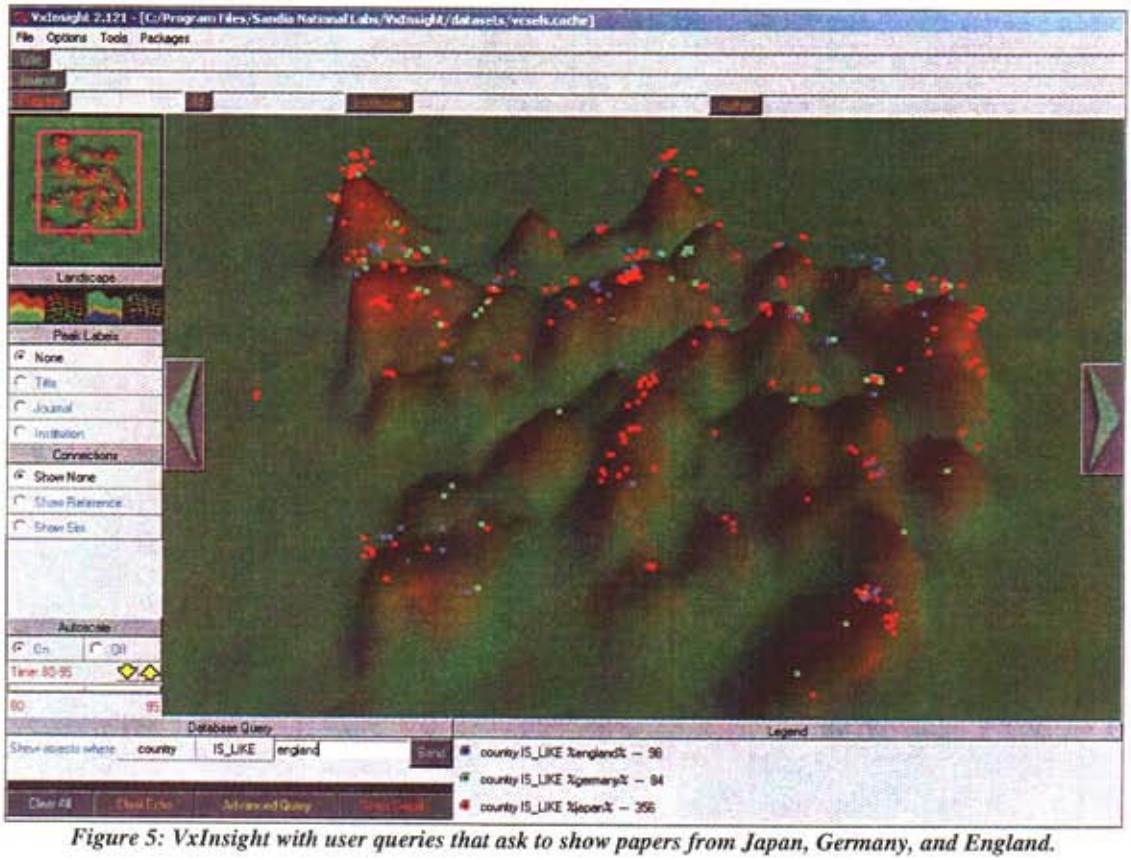


Figure 5: VxInsight with user queries that ask to show papers from Japan, Germany, and England.

One of the most powerful features of VxInsight is the ability to query a database. For instance, the analyst can use the SQL interface to make a database query about all the papers that were published in various countries. These queries would show up as colored dots lying on the terrain directly above the associated papers as can be seen in Figure 5. Now one can quickly pin-point the concentrations of these countries in the research area being explored. When used with the time slider, visual specifications of matching queries can be used to track the ebb and flow of competitive advantage as represented in the published record

Visual Display Techniques

One of the guiding principles behind VxInsight was the need to provide multiple resolutions of the data set. Providing seamless traversal through viewing resolutions allows greater comprehension of the data on the whole. The principle drawback to the multi-resolution concept is its computational cost. Allowing the user to have all these different views involves either dynamically computing the views on the fly or pre-processing the data and storing various resolutions to be called up at run-time.

With visualization of large datasets in mind, VxInsight has chosen the dynamic route for computing multi-resolution views. When exploring large information spaces, much of the memory is allocated to holding the data, leaving little for the visualization. Also, depending on the dataset, there may be many orders of magnitude difference from the highest to the lowest view scales, so storing all scales would be prohibitively memory intensive. Instead, VxInsight computes and provides *detail on demand* (Shneiderman, 1996).

Detail on demand refers to the postponing of detail computation until it is requested by the user. First, an overview of the entire collection is presented. Given an overview, the analyst can now zoom in on areas of interest. Only these areas are recomputed at higher resolution.

Software tools must be available to a large customer base. In the development of VxInsight steps were taken to minimize platform dependencies and exclusions. The VxInsight environment may appear to require extensive graphics hardware, but in fact uses polygons quite sparsely. To dynamically generate the terrain at various resolutions we overlay a regular grid on the currently viewable data objects. All objects fall into one of the grid bins, and then these bins are used in both the terrain height computation and labeling algorithm. By using this approach we find a good compromise between modest resolution and interactive rates for dynamic terrain generation. The regular grid technique requires only a constant number of polygons no matter what the resolution level. To be precise, the terrain consists of 3200 triangles, all t-striped to improve performance. This means that a modest PC (333 MHz) with an OpenGL graphics card is quite capable of running VxInsight.

Dynamically Configured Menus

When a dataset is accessed, a simple descriptor file is consulted. This descriptor file allows the analyst to customize the menus for a particular research focus. For example, when a dataset of web pages is being viewed, the informational display may have text widgets labeled 'Title' and 'Web Address'. The peak labels can be set to either 'Title' or 'Company' and the 'Connections' menu allows us to show 'Web Links'. Now by simply opening a dataset based on financial dealings, the entire informational display and menuing options transform. The information display is now composed of text widgets reading 'Transaction Type', 'Amount', and 'Date'. The peak label options are now 'Buyer', 'Seller','Institution', and the 'Connections' menu allows us to show either 'Country' or 'Bank' links in the data. The SQL query options will also be automatically updated. The SQL query window allows boolean conjunctions and negations using the table names of the current database. These are immediately presented to the analyst, which is very useful when initially learning to use a new database, or when you come back to a database after an extended period of time and have forgotten the field names.

VI. Analysis methods

An overview of steps used in doing analysis with the VxInsight tool is given here. Analysis can be done at multiple levels of resolution, giving different information at each level.

Macro level

After a data set has been loaded into VxInsight, the first level of analysis is done at the macro level, meaning at the level where the entire data set is visible. One sequence of steps that can be used is detailed here. Of course, the analyst can alter the order of the sequence or repeat many of the steps with different queries, etc.

- First, note the size, shape, and position of the peaks. The height of a mountain reflects the density of the objects on the grid underneath the peak. Very tall but narrow peaks indicate a very tightly focused technology group with many items. Reasonably tall, but broader peaks indicate a focused technology area, but with perhaps more distinctions or subareas than in the narrow peak. Short peaks indicate a technology concentration with less activity.

- Toggle the landscape to wireframe mode (wireframe is the mode shown in Figure 2). This reveals the positions of the individual articles on the plane beneath the peaks. Note any bridging structures (i.e. articles that fall between major clusters). These often indicate research bridging two major technologies.
- Turn on peak labels to show the content of the major peaks. Peak labels will reflect dominant content in each peak. Toggle labels between different labels fields (e.g. title, terms) to get additional information about the content of each peak. Titles, keywords or terms, and abstracts contain different levels of information (Kostoff et al, 2000), so it is important to see the content from each. Note the content of the peaks from the labels and become familiar with the layout of the landscape.
- Note the positions of peaks relative to each other. Adjacent peaks indicate related technology areas. Ask questions such as: Does the placement of technology groups make sense on the macro level? Do adjacent peaks tend to reflect my intuition about adjacent technologies? If the answers to these questions are "no", go to the micro level to determine why.
- Turn on the links between objects (in wireframe mode). Note the major relationships (based on common content) between major peaks, as indicated by a high concentration of blue lines between objects in the two clusters.
- It is informative to query the underlying database and see where particular instances lie. There is often overlap between technologies. Query abstracts and terms for main areas of interest (e.g. fire sensor, smoke sensor, CO sensor, smart sensor, false alarms, IR sensor). Note the positions of query results on the landscape (as colored dots). Do the concentrations of query results match what you noted from the labels? Are the results of any one query found in just one cluster, or more? Are there small groups of outliers in the query results? These are questions that can be answered when zooming into the micro levels.
- Narrow the time slider to a one or two year window. Move the slider to the earliest time and view the results (this narrows the displayed landscape to the one or two year time of interest). Slide the slider through the years, and note changes in peak sizes, shapes, and labels over time. Are there clusters that were small and then grew larger each year as time progressed? Are there clusters that grew smaller each year? Are there any ridges that grow up between two peaks? Note any observations.
- Toggle between the colored mountain and wireframe modes as desired at any step in the process.

Micro levels

- The VxInsight zoom function allows for continuous (rather than predetermined discrete) resolution zooming into or out of the landscape. Thus, analysis can be done at many different micro levels.
- Zoom (or drill) into a peak of interest. Note if the high level cluster breaks into smaller but discrete clusters. Note the peak labels for content as they indicate smaller subsets within a technology area, and their interrelationships. Do the relationships of these smaller clusters make sense? If the labels to not give enough information, <shift-click> on one of the objects to get the database information (full title, list of terms, author, source, abstract, etc.) for that object into the display at the top of the VxInsight window. Note the information, and click on other objects in the same region to accumulate information on the cluster.
- Zoom into ridges to determine the nature of the bridging between two technology areas. Is this of interest to you?
- Use the time slider at this level as explained above. What are the most recent activities? Do you percieve them as cutting edge?
- Zoom into query results that lie outside the main cluster of query results. Unexpected areas with query matches can be of great interest. Zoom into each individually and <shift-click> to view its database information. Click on surrounding documents. Does it look like the document matching the query result is a newer application of a technology that has been maturing for another purpose (e.g. a gas sensor technology that is now being applied to fire sensing)? These may indicate technology fronts, or disruptive types of technologies that will become mature for your application in the future.
- Throughout this process, you will find articles that you want to read. To read the full abstract for an article, open the MS Access file, and go to the 'forms' tab. Open the 'ViewRecords' form and scroll through until you find the ID number of the article of interest. It can also be beneficial to scan all the articles in a particular cluster. Two computer screens are helpful in this regard, but a single computer screen can be used.
- Analysis should be done, if possible, with the assistance of a subject matter expert. Someone who has a working understanding of the field can immediately see and confirm relationships that he/she knows exists. The expert can also more easily come to an understanding of why relationships that were unknown to him/her are there, and can more easily project what may come of those relationships and technologies in the future.

VII. Applications

Use of any analysis tool typically starts from one of two points. In the first case, one has questions to answer. In the second case, one has data to analyze and understand. We have used VxInsight in both types of situations and show illustrative examples of each here.

The most obvious applications of VxInsight are in analyzing literature and/or patent data to get a detailed mapping of a specific technology. This may occur in the information gathering stages of a project, in road mapping or forecasting events, or as a competitive intelligence activity. Questions that one may ask include "What are the various specialties in the field and how do

they relate to one another?" "What are the research fronts?" "Who is doing the work?" "Are there potential connections between two apparently unrelated technologies of interest that have not yet been made?" "Has someone else already done what I want to do?" "Should I develop or license technology?" Information that will aid in answering these types of questions can be gleaned from maps of science and technology based on literature or patent data.

Microsystems

We have used VxInsight to answer specific questions regarding many different technologies of interest to Sandia National Laboratories. For example, Sandia has developed expertise in microsystems engineering over the years as a part of its stockpile stewardship mission, and wanted to survey the field to identify potential collaborators for specific projects. Relevant technical terms were taken from a survey article on microsystems (Picraux & McWhorter, 1998) and used to query titles and keywords from the Science Citation Index (SCI). A total of 20,923 articles from the years 1990-1999 matched the query terms. Of these, 13,433 articles were connected to at least one other article in the set by citation. A similarity measure was generated from the citation structure using the combined linkage method of Small (1997). Ordination was done using the VxOrd force-directed placement algorithm, resulting in the VxInsight map shown in Figure 9. Articles written by Sandia authors are shown as yellow dots on the landscape.

The first step in analysis of the microsystems map was to understand the lay of the land. Navigation and query revealed that four main specialties populated the landscape: quantum dots and wires; nano-scale technologies; microsystems, MEMS, and other micro-components; and monolayer technology. Approximately 90% of the articles in the map were directly related to one or more of these four specialties. High-level analysis of trends shown by the microsystems map was identified using the time-sliding function of VxInsight. The landscape view was limited to the articles within a two-year period of time. Consecutive two-year periods were viewed, and resulting shifts in the peaks and valleys within the landscape were noted. For example, Figure 10 shows the two-year periods of (a) 1994-95 and (b) 1998-99. The peak at the top of the landscape has grown in size, indicating more publishing activity in the later years. Additionally, the peak label has changed from 'GaAs / Self-assembled' to 'InAs / Self-assembled', indicating a shift in the materials used in the quantum dot technology from GaAs to InAs. Query results shown as colored dots over the landscape support this interpretation, and also show a shift from 'grown' (green query results in Figure 10a) to 'self-assembled' (blue and yellow query results in Figure 10b) quantum dots.

A more detailed analysis was also conducted. Queries were made to identify the institutions doing the most work in the microsystems field. More detailed queries revealed the number of articles from each institution in the four main specialties, thus indicating relative strengths of various institutions. Accumulation of these results revealed differences between types of institutions as well. US universities as a whole divided their research nearly equally between the microsystem, quantum dots and wires, and monolayer areas, and did slightly less work in the nano-technology area. By contrast, US industry was heavily focused on quantum technologies, while US government laboratories focused on quantum and monolayer research. Research at foreign (Japanese and European) industry was even more heavily focused on quantum technologies than was research at US firms. While nano-technology work was less prevalent than the other three specialties, it had the largest overlap with the other specialties than any of the other three areas.

A final analysis was done to identify those institutions whose expertise closely matches Sandia's in specific areas. One way to find these types of synergies is to drill down into small areas within the landscape and follow citation links to and from Sandia papers (see Figure 11). This helps identify not only institutions, but specific researchers as potential collaborators, and also shows relevant prior or existing collaborations that have resulted in publication.

Analyses such as these where the data comes from the SCI have obvious benefits and limitations. A primary benefit is that the citation structure provides a defensible basis for mapping the structure of science, and lends itself well to detailed analysis. However, the SCI does not provide complete coverage of all literature that may be relevant to a topic of interest. Specifically, it includes few proceedings or conference papers. Inclusion of these types of articles from other sources should allow more accurate answers to questions such as those listed above.

Text mapping

Another area of interest to Sandia is the space within which it operates as compared to its sister institutions, Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL). These three laboratories present work at many conferences and also publish many government reports. Thus, it was felt that the SCI alone would not provide the coverage we desired.

Queries were made to five bibliographic databases: SCI, Cambridge Scientific Abstracts (CSA), Engineering Index (EI), INSPEC, and Medline, to extract all articles authored by SNL, LANL, and LLNL during the time period 1997-99. A common format was used to combine information from each of the original sources (see Table 1). A total of 28,329 articles were extracted from the five database sources, of which 16,939 unique articles were kept for analysis. Duplicate articles from the merged data source were eliminated by identification of identical (or nearly identical) abstracts. Extracts from the various sources were added

(and duplicates identified) in the order listed in Table 2. The increase in the fraction of duplicates with each data source is largely due to the order in which the data sources were added. Table 2 shows that the total number of articles was more than double the number of articles that were available from the SCI alone. Thus, our goal of obtaining more complete coverage of the technology space seems to have been met.

Combined	SCI	INSPEC, EI	CSA	Medline
ID	ID	ID	AN	UI
Abstract	Abstract	AB	AB	AB
Org	Org, City, State, Zip, Country	IN	AF	AD
Source	Journal, Vol, Year, Page	JN	SO	SO
Year	Year	parse from PB	PY	DP
Type	Type	DT	PT	PT
Title	Title	TI	TI	TI
Author	Author	AU	AU	AU
Terms	Keyword	DE	DE	MH
	Citation structure			

Table 1: Information source formats

Source	# Extracted	# Kept	# Duplicates	% Duplicates
SCI	8325	8325		
CSA	5858	3451	2407	41.1%
EI	4145	1916	2229	53.8%
INSPEC	9534	3172	6362	66.7%
Medline	467	75	392	83.9%
TOTAL	28329	16939	11390	40.2%

Table 2: Number of articles from each source

The choice of a field on which to base similarity is very important, and should be done using a field which has rich and common content across all contributing data sources. Citation structure is not common to all sources and cannot be used. A short list of fields containing content potentially suitable for calculation of similarity includes titles, terms, and abstracts. Titles are often too short to provide any comparative content. Different database vendors assign different sets of descriptive terms to similar articles, thus making terms a dubious choice for comparison (Qin, 2000). Thus, we have chosen to base similarity on the textual content in abstracts.

Natural language processing (NLP)

To accomplish the next two steps, elimination of duplicates and generation of a similarity matrix, we have used commercially available NLP software. Many different software packages using NLP techniques are available. A recent survey of these packages to choose the best for our purposes has not been performed. A package was selected for which we already owned a license. This package is RetrievalWare by Excalibur Technologies. It has done a very good job of text indexing and scoring for us, but there are likely other software products that could provide the same function.

RetrievalWare has indexing, search, and retrieval capabilities using an advanced natural language processing engine, and works on libraries of text documents. RetrievalWare has a Query-by-Example (QBE) search function, which allows the user to use any document as the query. The software then looks at all other documents in the library, finds those documents most like the query document based on words, word locations, word frequencies, and potential multiple meanings, and returns a ranked set of documents, each with a ranking score.

We use RetrievalWare for two steps – 1) identification of duplicate records, and 2) calculation of similarities for ordination. The process detailed here is used for each of the two steps:

- The abstracts for each of the articles are placed in separate text files in a directory. This constitutes the abstract library.
- The abstract library is then indexed using the RetrievalWare indexing function. This step must be performed prior to any search and retrieval functions.
- Command line scripts run RetrievalWare to search for a single abstract by ID, select that abstract, and then run the QBE search. This script is repeated for each abstract in the library, resulting in 'standard output' of the following form.

```
QUERY: A1990GE49500003
  1.( 47) A1990GE49500003
  2.( 30) A1999168MR00018
  3.( 29) A1997WG44400001
```

- The first ranked entry in this output is to the query document itself. The remaining entries are to other documents, where the number in parentheses is the ranking value. The top 10 ranked entries (excluding the self-reference) were kept for each article query.
- A program was written to read the output files and parse the ranking information into a similarity file with format

A1990GE49500003	A1999168MR00018	30
A1990GE49500003	A1997WG44400001	29

- The similarity pairs and values were evaluated using database functions for any occurrences where the transpose of a pair was also in the file, e.g. (A, B, sim_{ab}) (B, A, sim_{ba}). In these cases, the similarities were averaged as (A, B, $[sim_{ab} + sim_{ba}]/2$) and the (B, A, sim_{ba}) pair was eliminated from the similarity file. This results in a single-directional similarity matrix (i.e. non-zero values populate only half the matrix), which is the preferred format for the VxInsight clustering algorithm.

There are advantages and disadvantages to using a fixed number of ranked similarities as a similarity measure, rather than using every ranked similarity that could be generated. Advantages include 1) each object (article) will be linked to a minimum of ten other articles, thus each article should be found in a cluster, and 2) there is not excessive cross-linking, thus the ordination should produce many clusters. One disadvantage is that this approach places an arbitrary cut-off on what should be included in the similarity matrix, thus ignoring some potentially valuable information.

Elimination of duplicate records and calculation of similarity values

Duplicate records were identified by comparing abstracts from all records using RetrievalWare, and eliminating those that were identical or nearly identical. Previous study has indicated that nearly all duplicates either had query scores of 40 or more, or had scores within 4 points of the self-query score. Thus, the following filtering criteria were established to remove duplicates.

$$(RANK \geq 40) \text{ or } (RANK - SELF_RANK \leq 4)$$

Records from the EI Compendex were preferentially kept in cases where the EI and INSPEC records were duplicated, since the raw data from the EI queries can be transferred to LMIS. A full list of duplicates records was compiled in this manner, and these records were eliminated from the abstract library.

The RetrievalWare process as noted above was run again, from re-indexing of the library through execution of the QBE scripts to calculate similarity values for the reduced (after elimination of duplicates) data set. The top 10 similarity values for each article are kept for inclusion in the similarity matrix. Ordination was done using the VxOrd force-directed placement algorithm, resulting in the VxInsight map shown in Figure 12.

Analysis of three-lab landscape

Analysis of this three-lab landscape was done using techniques similar to those used on the microsystems landscape. Noyons, Moed & Luwel (1999) raise the issue that journals and conferences may cover different topics. Our analysis of the three-lab landscape shows that although articles, conference papers, and government publications (green, orange, and blue dots, respectively in Figure 12) coexist in much of the map space, there are pockets in the landscape where one type of publication predominates. This is particularly true near the bottom of the map where there are concentrations of orange and blue dots, corresponding to the areas such as high-energy physics and work with radioactive substances. Although it may be argued that the DOE laboratories specialize in these areas to some degree, it suggests that specialization between journals and conferences may occur in other fields as well.

In addition to mapping literature spaces, we have also used this text mapping technique to produce landscapes of Sandia discretionary R&D activities using proposals submitted by individual researchers. Analysis of these data allows Sandia to identify potential collaborations within the company (which are not always obvious within a large research institution), and to track the shift in internal R&D spending in different technical areas from year-to-year. Inclusion of corporate goals and line-of-sight descriptions in the landscape allow to correlate and align investment with corporate goals.

Journal mapping

Collaborative efforts between Sandia and other institutions (particularly universities) have been enhanced through the generation of maps of science based on the relationships between journals. One such effort was to create a landscape based on citations between the 4.7 million articles from the 3000 journals in the physical sciences subset of the SCI from 1981-1996.

The similarity between journals was defined as (A cites B) + (B cites A) where A and B are any two journals. Normalization, such as that in Jaccard-type similarities, was not used. This naturally provides higher weights to interactions between the larger journals, thus producing a map where the large journals form clusters, and the smaller journals agglomerate to the large journals they are most like. Ordination of journals was done using the VxOrd algorithm. The resulting landscape (see Figure 13) shows

that Physics occupies a central position relative to other disciplines. Engineering disciplines, mathematics, computing, and materials science all surround physics with strong links to physics journals. Various chemistry disciplines occupy the ridge on the right edge of the map and have strong connections back to physics as well as to materials and geology. At the macro-level this map has much in common with the physical sciences portion of the map generated by Bassecoulard & Zitt (1999).

A close-up view of a cluster in the center of the astrophysics peak (Fig. 13b) shows strong relationships (blue lines) between major journals in that field. It is interesting that *Nature* appears in this cluster, although it is a multidisciplinary journal. In this map where life sciences are not included, *Nature's* strongest links are to well known astrophysics journals. Difficulties associated with clustering multidisciplinary journals have been enumerated by other authors (Bassecoulard & Zitt, 1999).

Sandia has used maps of science to compare the output of many research institutions in different fields of science. This allows us to know the relative emphases placed by a potential strategic partner in different fields. Both overlapping emphases and complementary emphases are important information when pursuing a strategic partnership.

Genome mapping

The VxInsight™ data-mining tool has been extended to visualize data in genomic databases, which are collected from microarray experiments (Spellman, *et al.* 1998; Lander, 1999; Schena, 2000). It presents a novel and intuitive way to cluster and view those gene expressions and provides a way to access other web-based resources while maintaining the context associated with VxInsight clusterings.

Similar genes are clustered together spatially and represented with the 3D terrain map, where the large mountains represent a large clusters of similar genes and smaller hills represent clusters with fewer genes, and the more similar clusters are also sited more closely together, see Figure 6. The terrain metaphor is very intuitive; most biologists are able to work with the user interface immediately. The analogy to real terrain allows one to memorize the "landscape," which makes subsequent explorations easier, making the analyst more productive.

The highest-level view of whole genome-level expressions is useful, but VxInsight provides other critical data analysis techniques. First, one can zoom into the terrain to see greater detail (the finer clustering within larger clusters), and second, one can easily search expression and annotation databases with SQL queries. Rather than getting a long list of textual responses to these queries, the results are displayed within the context of the visual landscape by highlighting the genes matching the query as shown in Figure 7. The visual context can be as important as the result of the query, so in combination they are powerful tools for exploratory data analysis. Finally, VxInsight provides a simple a way to select genes and immediately link them to web-based resources through a web browser like Netscape Communicator, see Figure 8. This linkage is particularly important because an analysis typically involves understanding the local experiments within the context of the broader scientific/medical knowledge recorded in the literature and on the web (for example through Medline and in the databases maintained by Proteome Inc.).

VxInsight has been used to analyze microarray data from *Saccharomyces cerevisiae* and *Caenorhabditis elegans* as well as human gene expressions. In each case, a battery of microarray experiments resulted in a series of expression levels for each gene. To determine similarities between genes a correlation coefficient is computed for each possible pairing of the genes. Typically, only the twenty highest correlations for any gene are used to ordinate the VxInsight landscape. Various normalizations are used to enable better clustering. For example, the data from all the experiments is typically normalized by reference to each gene's expression in a baseline, or control experiment. Even with this normalization, the distribution of expressions will, typically, be heavily skewed to the right. As suggested by Tukey and Mosteller [Mosteller and Tukey, 1977], these distributions are log transformed to give a more nearly normal distribution, before computing the correlation coefficients. The top twenty expressions are sufficient for computing the ordinations, but selecting pairs of extreme similarity for the "strong connections" arcs is more problematic.

Being able to see arcs between the genes with the strongest similarities is quite appealing and could be as simple as just showing the top N, say 100, arcs. However, it is more satisfying to be able to say something about the statistical significance of those arcs. Certainly, one would like to be able to give a P-value indicating confidence that the arc was not connecting genes whose true correlation (rho) was zero. But that is a pretty weak statement, especially when the experiment involves O(10,000) genes and O(100) to O(1000) microarrays, because any tiny variation from zero will be statistically significant give that many observations. Further, with O(10,000) genes and O(10,000^2) comparisons, a huge P-value will be required to limit the number of false positives. Instead, we use an approximation due to Fisher (Ostle, 1963; Fisher,1921). When the underlying, true correlation rho_0 is zero, the estimated correlation, r, has a t-distribution, but when rho_0 is not zero the following transformation of the observed sample correlation will be approximately normally distributed with mean rho_0 and variance 1/(n-3):

$$Zr = 1/2 \ln((1+r)/(1-r)).$$

Given this formula and input from the user about some value of rho that must not be missed as significant, and the experimenter's willingness to risk a type-II error given the true rho, it is possible to compute a critical threshold for r, see Figure 7. This approach successfully identifies a small number of very significantly correlated gene expressions.

We have investigated the application of VxInsight to genomics for less than a year and have found it to be useful in practice. In the course of these investigations, we have identified new research questions that need to be carefully studied. For example, it is know that Pearson's correlation coefficient is not robust. That is, non-gaussian noise and/or the presence of a few outliers can have undue influence on its value. Various robust, non-parametric measures of correlation have been proposed [ref Cox], including the percentage bend correlation coefficient, Rpb, which we are now investigating, as a replacement to Pearson's correlation coefficient, r.

Another critical question still to be investigated is how stable are these genomic clusters? That is, will the clusters change significantly if a small amount of new data is added, or if some random subset of the existing data is not included? More generally, the question of cluster distributions should be addressed and measures of central tendency and variance should be developed. Reporting a single cluster in the summary of a data analysis is as poor an approach as reporting just the experimental mean of a dataset without error bars or some indication of the variance and distribution of that experimental value. Powerful statistical methods are available to investigate the distributions that result from different clustering algorithms (Mosteller and Tukey, 1977; Cox 1997). However, these bootstrap methods are computationally challenging and most biologists have very limited access to the scale of computing required, so we hope to continue further investigations with our resources.

Finally, the question of information content in any given microarray experiment should be addressed. Certain experiments have more or less information than others, which can be seen by the variability of the genes' expression values. For example, an experiment that reported all genes having the same expression as they had in a control experiment imparts no new information. On the other hand, an experiment with many genes that widely vary from the control has more information than an experiment with only a few genes up or down regulated. The amount of information in any experiment should be used to weight the correlation computations to yield better estimates of which genes are truly related. In subsequent research we hope to address all of these issues in some depth.

VIII. Summary and Future Work

The following general and specific conclusions about the VxInsight tool and analysis environment have been reached as a result of this and other detailed studies using the tool:

- The non-expert who has achieved some proficiency using VxInsight can quickly discover general characteristics in a technology area.
- Expert interaction in the forecasting process is essential, both in terms of search term formulation, and on the analysis end. Interaction with bibliographic data using VxInsight can aid the expert in technology forecasting efforts.
- Initial query terms are very important to exclude articles unrelated to the subject of interest.
- When using the Science Citation Index, a similarity measure based on citations may not be suitable to generate a useful and meaningful landscape due to insufficient linkage between articles.
- Use of a similarity measure where the similarity value is the common number of keywords between any two articles works well for a single database.
- Use of a similarity measure where the similarity value is the common number of keywords does not work well for an article set built from two or more bibliographic sources due to inconsistent methods of generating keywords.
- A similarity measure based on common textual material in abstracts works very well on merged databases and produces useful and navigable landscapes.
- Data from multiple bibliographic and patent sources provide much better coverage of the intended subject area than does one source, and can be merged with good success to produce meaningful landscapes.
- VxInsight displays data in a way that allows potential disruptive technologies to be more easily spotted and evaluated. These technologies may be represented by single articles or a small group of articles that reference a particular technology, but are not found on the part of the landscape containing the main body of work for that technology (i.e. outliers, but still within the map).

In a larger context, visualization tools, such as VxInsight, are becoming crucial to science and technology management in the context of textual data mining (TDM). Technical experts are proving to be crucial in this process. The Naval Research Laboratory has sponsored research over many years aimed at using textual sources in their S&T management processing, including roadmapping and the identification of promising technologies. Quoting from Losiewicz (2000),

> "The FY98 experience showed conclusively that high-quality data mining requires the close involvement of technical domain expert(s) in information retrieval, phrase frequency and proximity analyses, and presentation. Multiple perspectives are often needed to detect data anomalies ... TDM [textual data mining] cannot realize its full potential in S&T management

if used only sporadically -- it must become an integral part of the S&T sponsor's business operations. Because of the learning curve, long-term involvement of experts with data mining experience in a particular topic area is desirable."

Development of a long-term capability for S&T management, roadmapping, and forecasting, while not an easy process, will prove to be increasingly valuable as we proceed in this information and technology-rich age in which we live and work.

IX. Figures

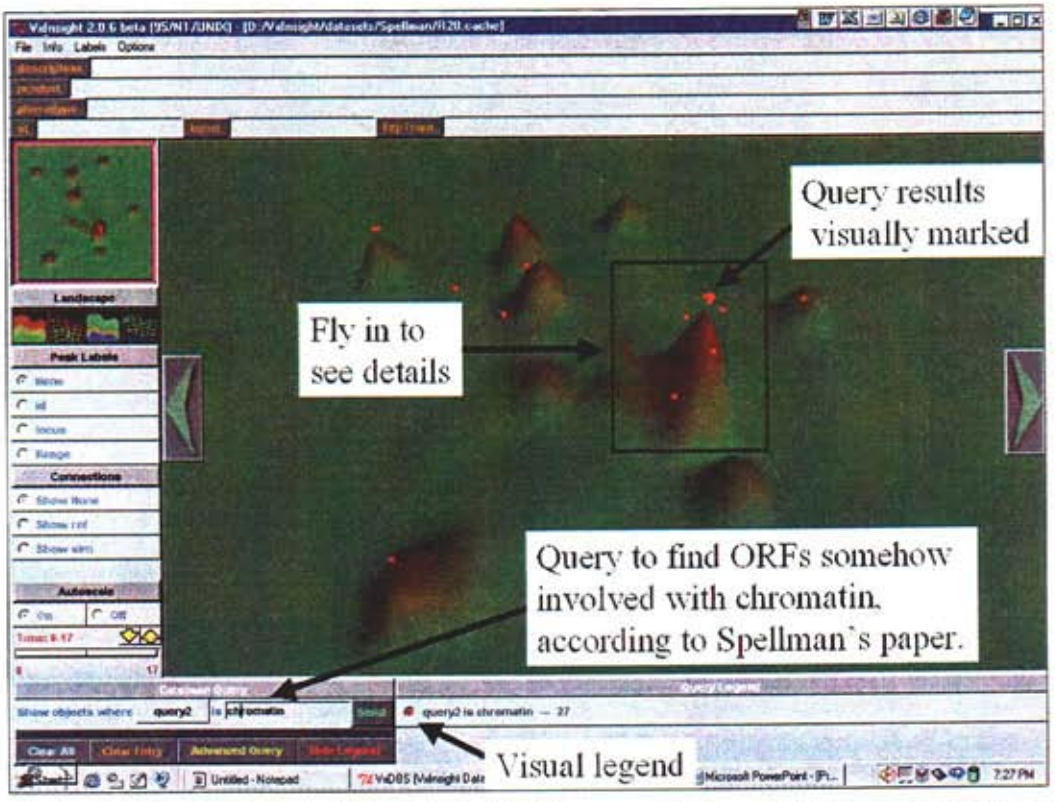


Figure 6: Yeast gene expression data visualized with VxInsight. Finer resolution is available by zooming into the clusters, and database searches can be entered using SQL. Here various genes associated with the production of chromatin are highlighted.

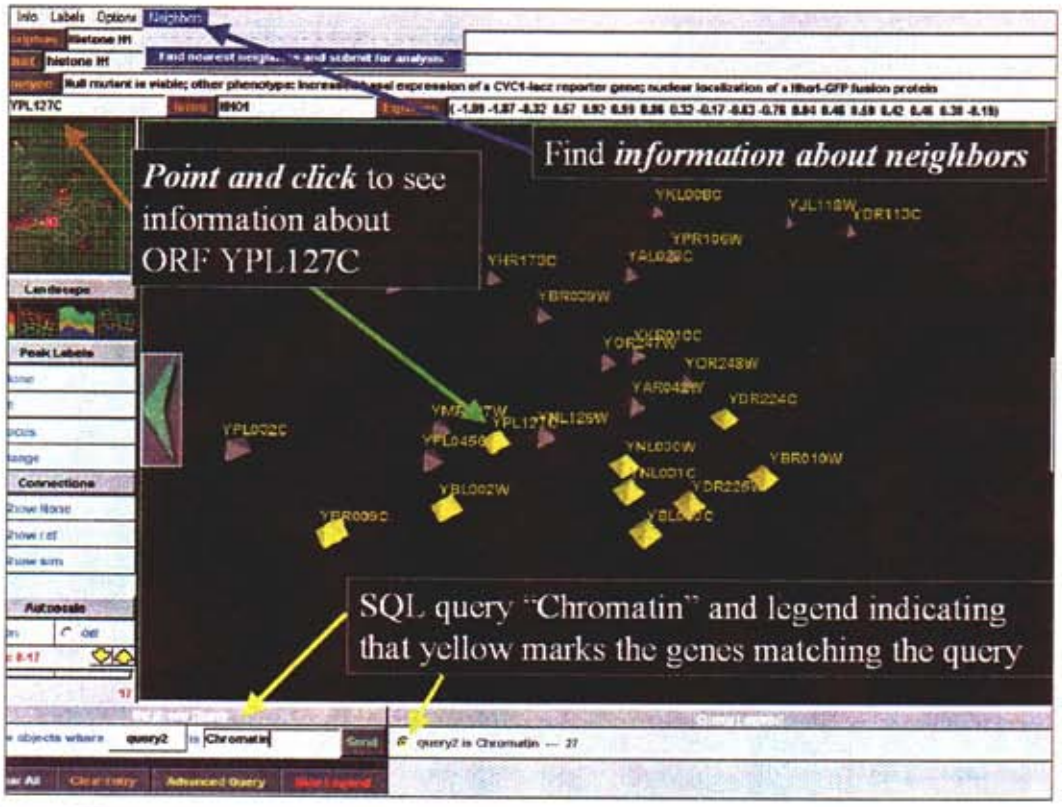


Figure 7: An SQL query asking for genes associated with chromatin is shown. Note that the results of the query are shown in yellow and in the context of their clusters.

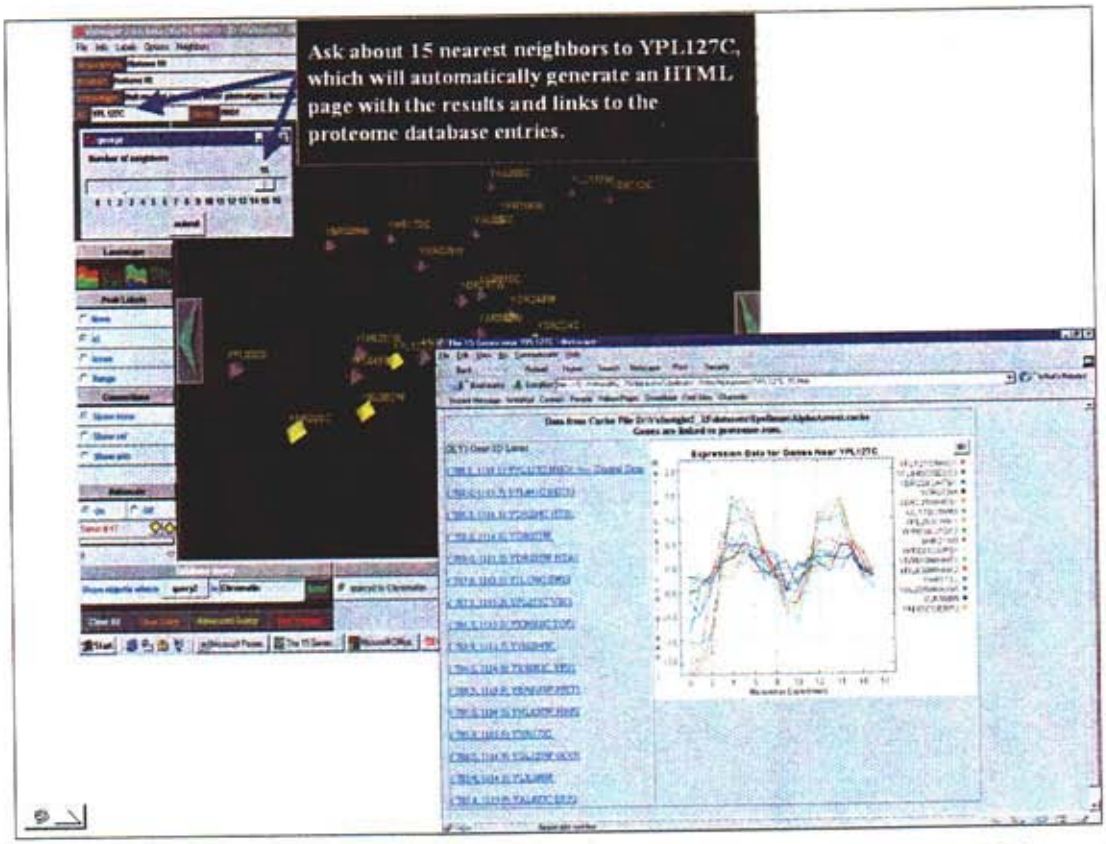


Figure 8: An example of using VxInsight to generate an HTML page with a gene expression plot and links to web-based resources describing each of the genes near YPL127C.

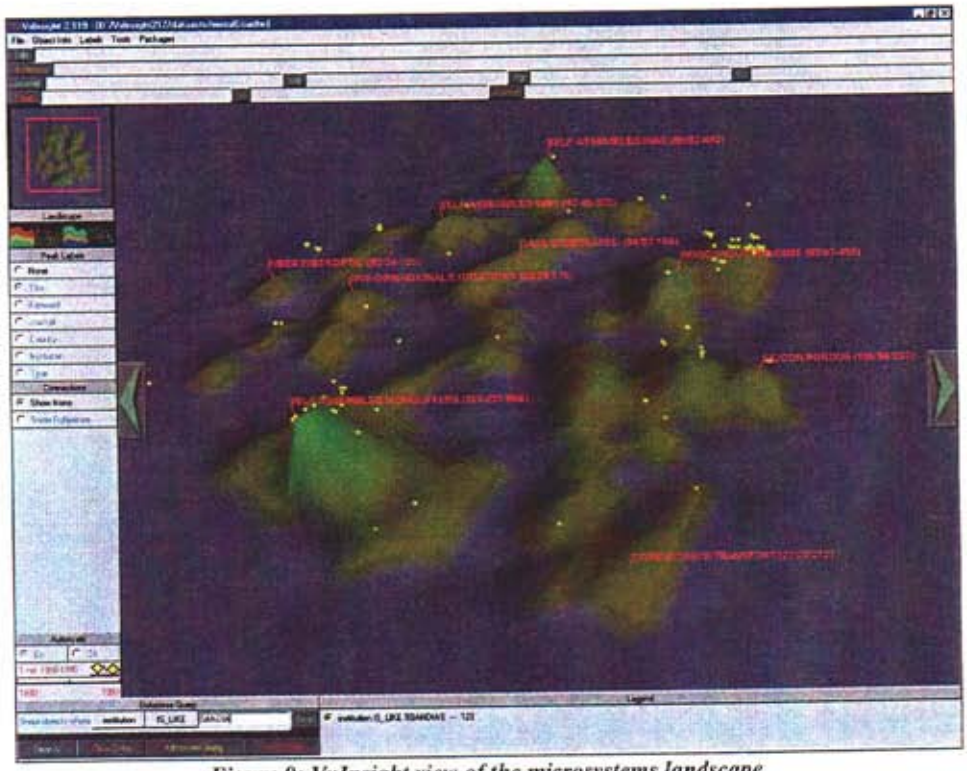


Figure 9: VxInsight view of the microsystems landscape.

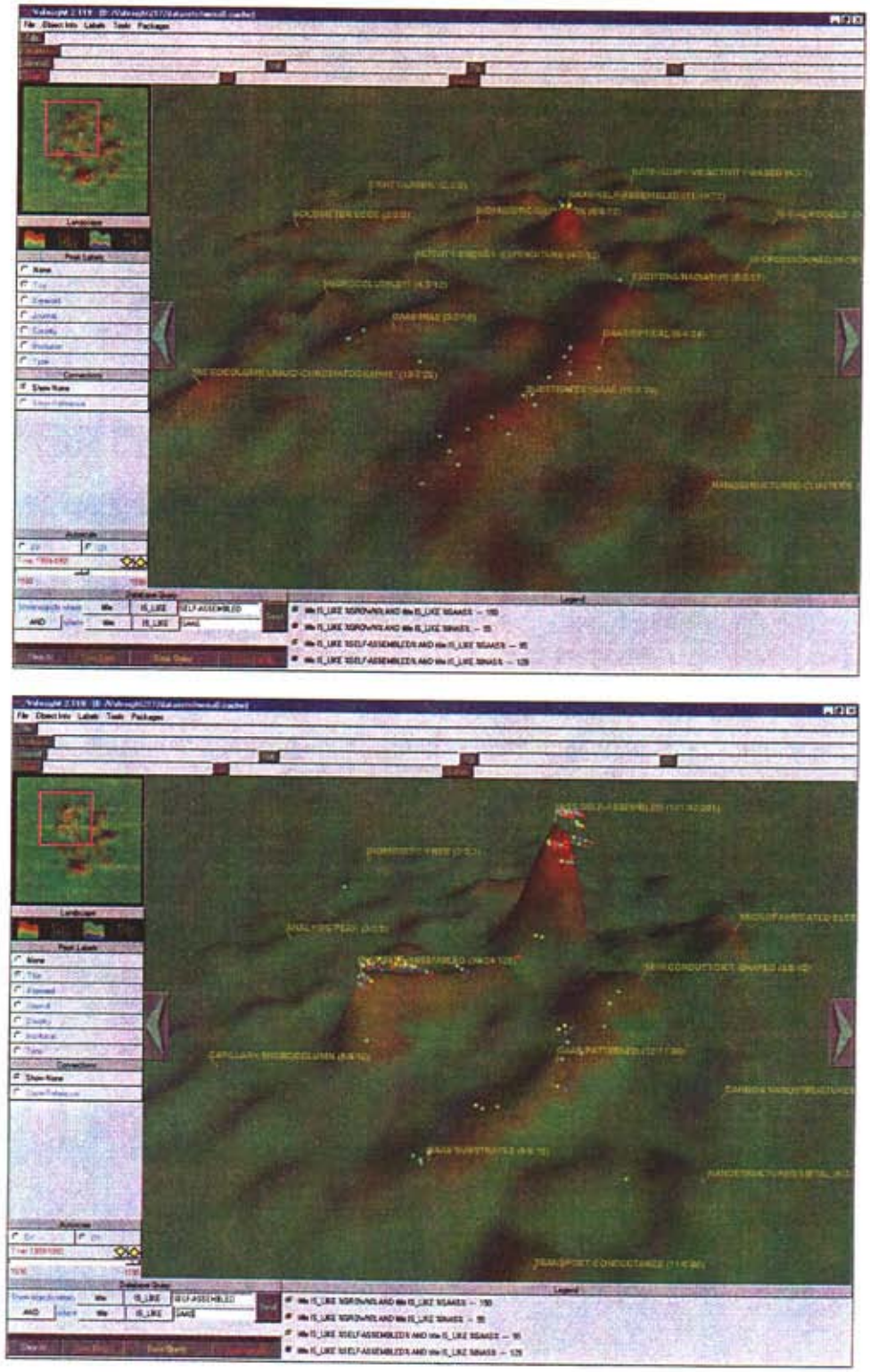


Figure 10. Microsystems landscape – two 2-year periods (94-95) and (98-99).

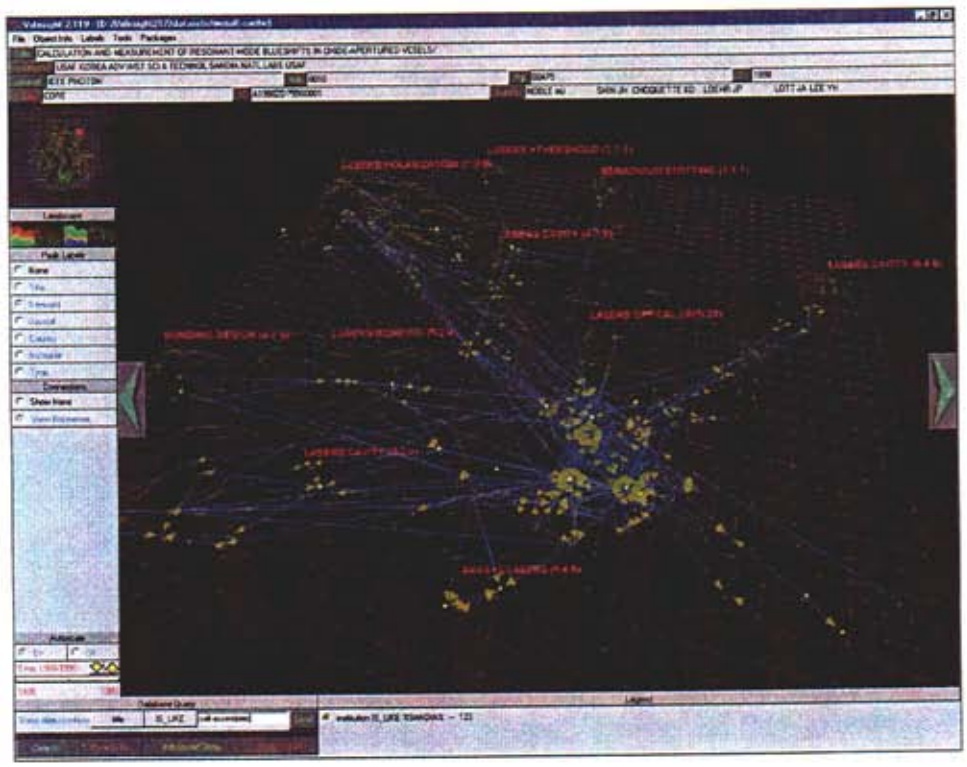


Figure 11: Detailed view of small portion of microsystems landscape.

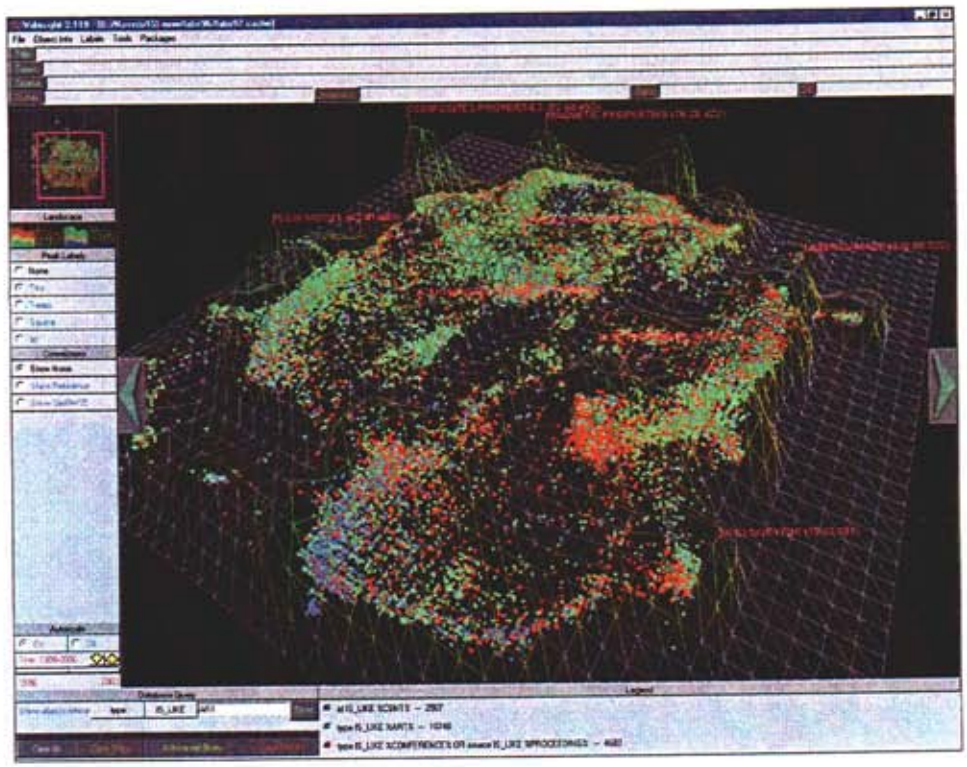


Figure 12: VxInsight view of the three-labs technology landscape.

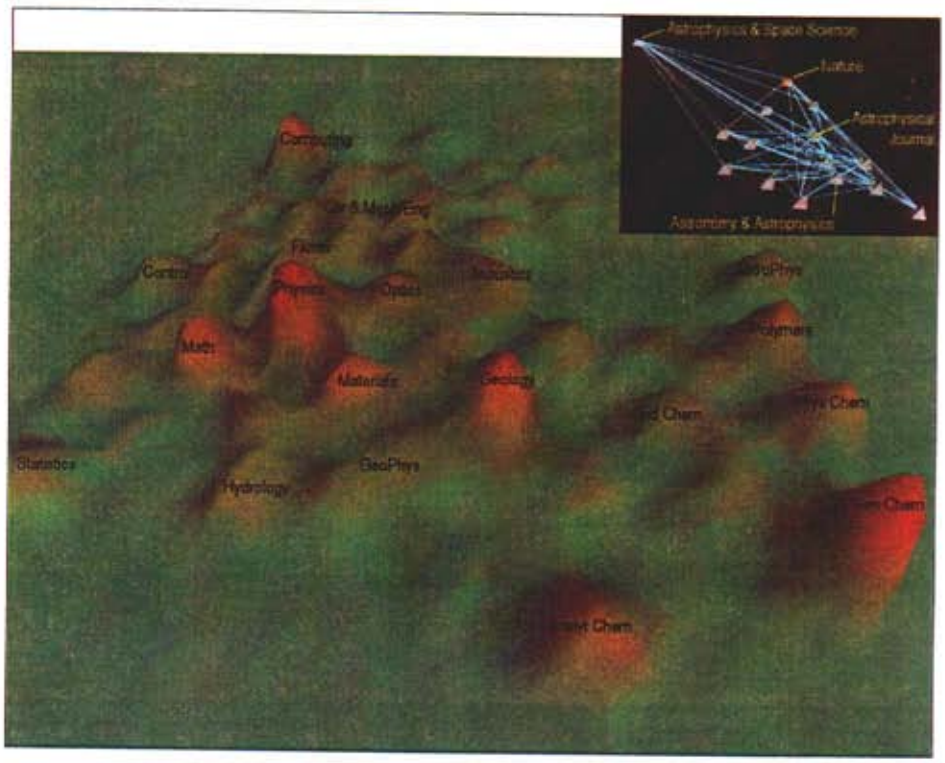


Figure 13: Landscape of the physical sciences displayed in VxInsight™. 13b) Detail on the Astrophysics cluster from 13a. Strong links between individual journals in the cluster are shown as blue lines.

X. References

Bassecoulard, E. & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of scientific journals. *Scientometrics*, 44, 323-245.

Beck, D. F., Boyack, K. W., Bray, O. H. & Siemens, W. D. (1999). Landscapes, games, and maps for technology planning. *CHEMTECH*, 29(6), 8-16.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35, 401-420.

Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E. & Wylie, B. N. (1998). Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems*, 11, 259-285.

Fox, K. L., Frieder, O., Knepper, M. M. & Snowberg, E. J. (1999). SENTINEL: A multiple engine information retrieval and visualization system. *Journal of the American Society for Information Science*, 50(7), 616-625.

Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. (1998). *Multi-faceted insight through interoperable visual information analysis paradigms*. Proceedings of IEEE Information Visualization '98, 137-144.

Honkela, T., Kaski, S., Kohonen, T. & Lagus, K. (1998). Self-organizing maps of very large document collections: Justification for the WEBSOM method. In I. Balderjahn, R. Mathar & M. Schader (Eds.) *Classification, Data Analysis, and Data Highways*. Berlin: Springer.

Jain, A.K., Murty, M.N. & Flynn, P.J. (1999). Data clustering: A review. *ACM Computing Surveys* 31(3), 264-323.

Kostoff, R. N., Eberhart, H. J. & Toothman, D. R. (1999). Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. *Journal of the American Society for Information Science*, 50(5), 427-447.

Leydesdorff, L. (1994). The generation of aggregated journal-journal citation maps on the basis of the CD-ROM version of the Science Citation Index. *Scientometrics*, 31, 59-84.

Losiewicz, P., Oard, D.W. & Kostoff, R.N. (2000), Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, 15(2), 99-119, 2000.

McCain, K. W. (1998). Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field. *Scientometrics*, 41, 389-410.

Nederhof, A. J. & Van Wijk, E. (1997). Mapping the social and behavioral sciences world-wide: Use of maps in portfolio analysis of national research efforts. *Scientometrics*, 40, 273-276.

Noyons, E. C. M. & Van Raan, A. F. J. (1998). Advanced mapping of science and technology. *Scientometrics*, 41, 61-67.

Noyons, E. C. M., Moed, H. F. & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science*, 50(2), 115-131.

Picraux, S. T. & McWhorter, P. J. (1998). The broad sweep of integrated microsystems. *IEEE Spectrum*, 35(12), 24-33.

Qin, J. (2000). Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science*, 51(2), 166-180.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9), 799-813.

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, 38, 275-293.

Spasser, M. A. (1997). Mapping the terrain of pharmacy: Co-classification analysis of the International Pharmaceutical Abstracts database. *Scientometrics*, 39, 77-97.

Shneiderman, B (1996). The eyes have it: A task by data type taxonomy for information visualizations. In Proc. IEEE Symp. Visual Languages '96 IEEE, 1996.

Tufte, E.R (1992). The visual display of quantitative information Graphics Press, 1992.

Wise, J. A. (1999). The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13), 1224-1233.

Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1(4), 3.

Lander, E. S. (1999). Array of Hope, *Nature Genetics* 21(1) Supplement, 3-4.

Ostle, B. (1963). *Statistics In Research, Basic Concepts and Techniques For Research Workers*, The Iowa State University Press, Library of Congress number 63-7548.

Schena, M. (Ed.) (2000). *Microarray Biochip Technology*, Eaton Publishing, ISBN 1-8811299-37-6.

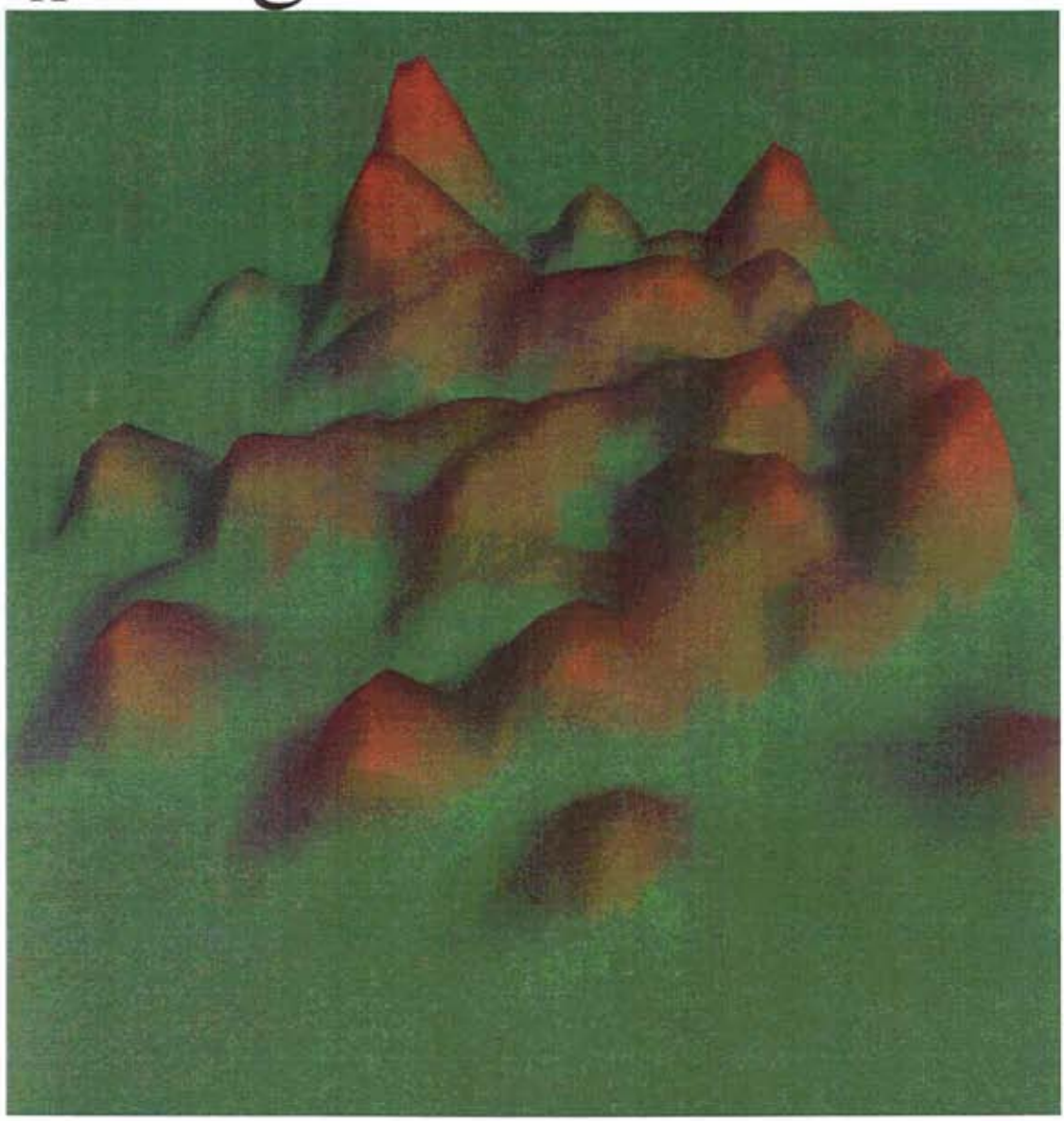
Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, 9, 3273-3297.

Mosteller, F. & Tukey, J. W. (1977). *Data Analysis and Regression, A Second Course In Statistics*, Addison-Wesley, ISBN 0-201-04854-X.

Wilcox, R. R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, ISBN 0-12-751545-3.

XI. Appendix

V_xInsight™



User's Guide

Version 2.x

Sandia National Laboratories

Table of Contents

INTRODUCTION **28**

INSTALLATION **28**

DISPLAY CONFIGURATION **28**

A. CONFIGURATION FILE NAMING CONVENTION 29

B. DATABASE CONFIGURATION FILE 29
*1. **Extension*** *29*
*2. **Entry Format*** *29*
*3. **Mandatory Information*** *30*
*4. **Special Information*** *30*

C. MENU CONFIGURATION FILE 30
*1. **Extension*** *30*
*2. **Menu Item Labels*** *31*
*3. **Entry Format*** *31*

D. STOPWORDS FILE 32
*1. **Extension*** *32*
*2. **Entry Format*** *32*

MENUS **33**

A. TOP MENU BAR 33
*1. **File Pulldown Menu*** *33*
*2. **Object Info Pulldown Menu*** *35*
*3. **Labels Pulldown Menu*** *36*
*4. **Tools Pulldown Menu*** *37*
*5. **Packages Pulldown Menu*** *37*

B. SIDE MENU BAR 38

C. BOTTOM MENU BAR 39
*1. **Database Query*** *39*
*2. **Legend*** *40*
*3. **Query Mode*** *41*

VISUAL DISPLAY AND NAVIGATION **43**

A. VISUAL DISPLAY 43
*1. **Terrain-View Mode*** *43*
*2. **Data Record Mode*** *45*

B. NAVIGATION 46
*1. **Zoom*** *46*
*2. **Rotate*** *46*

DATABASE INFORMATION RETRIEVAL **47**

A. DATABASE QUERY 47
B. SELECT DATA OBJECTS SUBSET 47
C. VISUAL DISPLAY OF QUERY RESULTS 50
*1. **Query Legend*** *50*

2. *Terrain-View Mode Display* *51*
3. *Data-Record Mode Display* *52*

GENERAL INFORMATION **53**

VXINSIGHT USER GUIDE

INTRODUCTION

In an effort to alleviate the information overload problem, the VxInsight application was developed as a data visualization tool. Using information obtained from a database, VxInsight generates a mountainous landscape that clusters data objects[1] by some user-defined similarity metric.

All of the information contained in the database is represented by the mountainous landscape, which the user can traverse at will. The user can zoom in and out of these mountain peaks to view specific information contained in the underlying data.

INSTALLATION

A self-extracting Windows NT Version executable will create a short cut to the VxInsight application on a user's desktop.

DISPLAY CONFIGURATION

There are three (3) configuration files needed by VxInsight for data visualization of information contained in one or more databases, i.e., the database configuration file, the VxInsight menu configuration file, and the stopwords file used by the label generation algorithm.

These files are related to a **cache** file used by VxInsight to display a landscape. The relationship is determined using the **filename prefix**. For example,

My_File.db (database configuration file)

My_File.config (menu configuration file)

My_File.stop (stopwords file)

are associated with the cache file, **My_File**.cache, and will be used during a VxInsight session displaying the "**My_File**" landscape. The above mentioned file formats are described below.

[1] "data object(s)" refer to a single record of information in a database

A. CONFIGURATION FILE NAMING CONVENTION

The naming convention for each of these files is as follows:

<filename>.<extension> where **<filename>** uses the naming conventions of the current system and should reflect the contents of the information to be displayed

where **<extension>** is specified for each file as described below

Because there can be several sets of these files available to a user, the **<filename>** portion for each in a given **"set"**[2] must match exactly. The **<extension>** determines what type of information the file contains.

For each installation of VxInsight, the configuration files are located in the **VxInsight** directory, **datasets**, i.e., **<base path>/VxInsight/datasets**. In the "**File**" menu of VxInsight, an "**Edit Files...**" button has been provided to allow a user to edit these configuration files.

B. DATABASE CONFIGURATION FILE

The database configuration file specifies the database(s), the tables in each database, and the fields in each table that are available for user access. Any combination of database(s), table(s) and fields can be used in this file as long as **VxInsight** can establish a connection(s) to the specified database(s).

1. EXTENSION

The database configuration file has the extension. "**.db**".

2. ENTRY FORMAT

Following is the format for each entry in the database file and a definition/description of each element.

<name> <hostname>::<database table name>::<field name> <field status>

where

<name> - a unique name used as an alias to represent a specific field in the database which will be used by VxInsight to access field information

<hostname> - refers to the name of the machine where the database is located

<database table name> - the name of the table in the specified database

<field name> - a field in a specified table

[2] "set" refers to one database, one configuration, and one stopword file for a specific data set

<field status> - determines if the field will be contained in the generated cache file

no – do not include this field in the local cache

cache – include this field in the local cache

connection – marks this field as one which will be used to generate relationship links

3. MANDATORY INFORMATION

The following alias and the corresponding field are mandatory in order for VxInsight to uniquely identify data objects. The actual alias name, **Id**, must be used to reference the database unique identifier field.

Id <hostname>::<database table name>::**<field name> cache**

where <field name> pertains to a unique identification field

4. SPECIAL INFORMATION

This special field alias allows the user to specify a range by which to narrow a window of information. If the user wishes to use the **Range** functionality, the actual alias name, **Range**, must be used to reference an integer field in the database. For instance, the range option could be an integer "year" field in the database.

Range <hostname>::<database table name>::**<field name> cache**

where <field name> pertains to an integer number field

C. MENU CONFIGURATION FILE

The menu items displayed during a **VxInsight** session are configured according to a user's specifications. Each **VxInsight** application installed has a "menu" configuration file that can be edited to meet the user's requirements for the specific data object information that is to be displayed.

1. EXTENSION

The menu configuration file has the extension. "**.config**".

2. MENU ITEM LABELS

Each alias/label, i.e., <name> as described in the entry format for the database configuration file, should be used when specifying the menu items to be displayed. Only those fields defined in the database file are accessible for menu items.

3. ENTRY FORMAT

a. Add a field to the data object information menu

Add_Small_Info_Field: This command creates a text field menu item as specified by the parameters. These fields are concatenated to create one menu item line for a maximum of eight per line.

Add_Medium_Info_Field: This command creates a text field menu item as specified by the parameters and is created one per menu item line.

Add_Large_Info_Field: This command creates a scrolling text field menu item as specified by the parameters. Depending on the size of the text, it is recommended that the height be kept to a maximum of 4 lines per menu item.

Format:

Add_<size>_Info_Field <name> <height> <width> <color>

where

<size> is "Small", "Medium" or "Large"

<name> is the label defining a specific field in the database

<height> is defined as number of lines

<width> is defined as number of characters per line

<color> determines the color of the label

b. Add a field for label generation

This command creates a toggle button menu item as specified by the parameters. Selecting one of these options during a VxInsight session will result in peak-label generation using the contents of the fields. It is highly recommended that fields selected for label generation are "locally cached" via the database file. Otherwise, the field information will have to be retrieved through a database query which will take a considerable amount of time thus slowing the landscape traversal as the labels are being generated. If a needed label field is not "locally cached", ask the database administrator to modify the database file. One or more labels can be designated as "peak" labels for the landscape.

Format:

Add_Label_Field <name>

where **<name>** is the label defining a specific field in the database whose contents will be used to generate a label on the landscape.

c. **Add a field defining a relationship:**

This command creates a toggle button menu item as specified by the parameters. A connection is an indication of some relationship between data objects in the database. It requires a generated "connection" table in the existing database showing the relationship between one or more sets of "ids" via some predetermined criteria. One or more fields can be designated as a relationship.

Format:

Add_Connection_Field <name>

where **<name>** is the label defining a specific field in the database whose contents will be used in determining a relationship between data objects on the landscape.

D. STOPWORDS FILE

The "stopwords" file is used by VxInsight in the peak label generation process. The words contained in this file are those, which a user determines, are not to be considered for label generation. This file can be modified as a user comes across irrelevant labels by selecting the "**Edit Files...**" button from the "**File**" menu and selecting the stopwords file from the file manager window. Once edited, the file should be saved.

1. EXTENSION

The stopwords file has the extension. "**.stop**".

2. ENTRY FORMAT

Each word is placed on a separate line in the file.

Upon startup, the "**title**" bar on the application window will contain the current version number of the VxInsight application.

Once a data set has been selected, the location and name of the data set cache file will be added to the "**title**" bar.

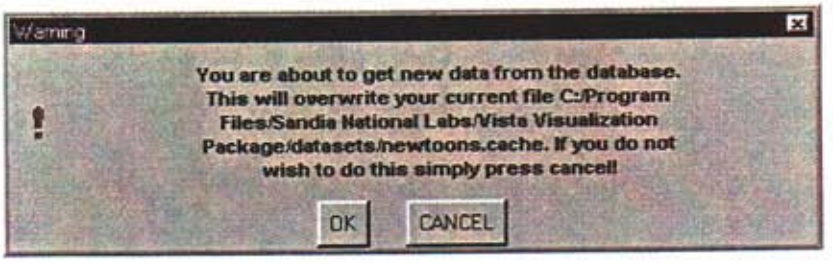
A. TOP MENU BAR

1. FILE PULLDOWN MENU

New ...(Create a new local cache file)

Allows user the potential of creating a new cache file. Upon selection of this option, the user will be presented with a file dialog box from which a specific data set database configuration file can be selected for creating the new local cache. The files contained in the dialog box will have a "**.db**" file extension. ". The user may browse directories but only files with the aforementioned selection will be included in the dialog box.

Once the configuration file has been accepted, VxInsight downloads the database data as specified by the configuration file. At this time, if a cache file with the same name exists, the user will be given the option to overwrite the file.

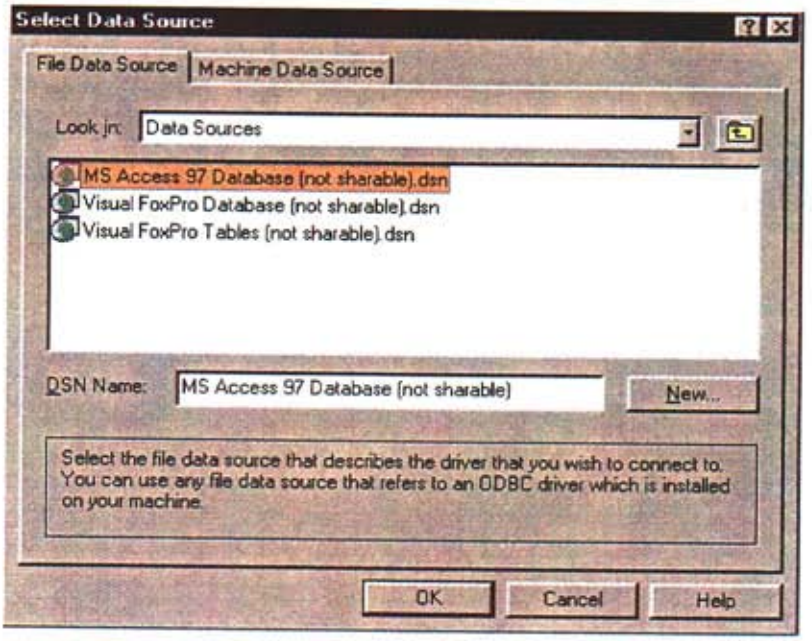


If the user chooses not to overwrite the file, the current process is halted and the user is returned to the previous process state.

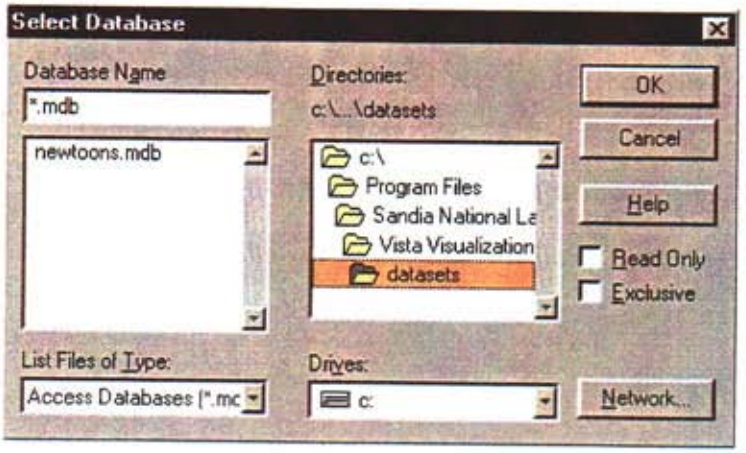
However, if the user chooses to overwrite the file or there was no existent file with the same name, the data download process will begin and the user will be presented with the following sequence of informational dialogs.

a. If a connection to the database has not been established prior to this sequence, the user will be able to connect to the database and continue.

The first dialog window allows the user to select from a variety of database sources.



Once a selection is made, the user will be able to browse for the database of interest.



b. With a connection established, VxInsight continues the **New…** process as follows.

Open … (View an existing data set)

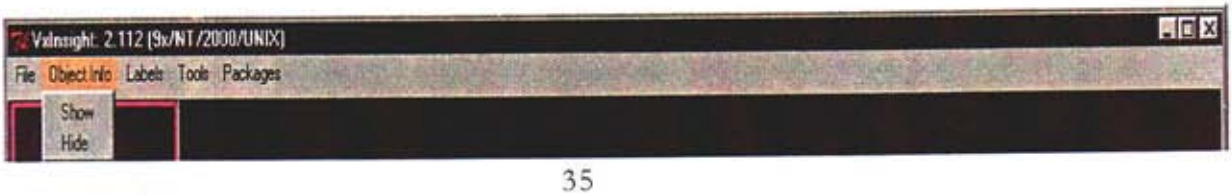
Allows user to open an existing landscape based on the specific data set selected by the user. Upon selection of this option, the user will be presented with a file dialog box from which a specific landscape data set selection can be made. The landscape cache files that are available for selection will have the extension **".cache"**. The user may browse directories but only files with the aforementioned selection will be included in the dialog box.

Save … (Save the current cache file)

Save As … (Save the current cache file with a new name)

Exit (Exit VxInsight)

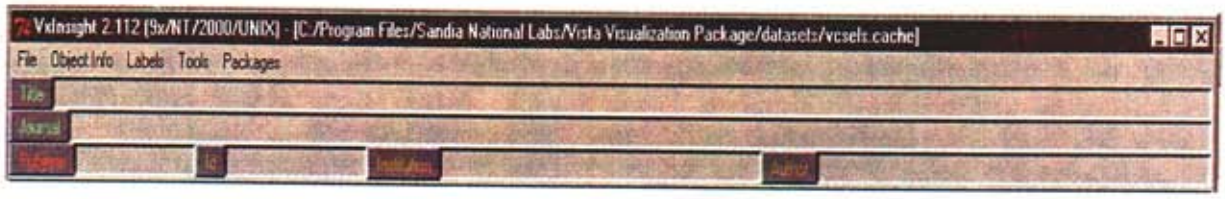
2. OBJECT INFO PULLDOWN MENU



Show… (Show the data object information menu)

Hide… (Hide the data object information menu)

The data object information menu is generated using the information in the **<data set>.config** file as described in the section, **Menu Configuration File,** referring to adding a field to the data object information menu. This menu will contain information about a selected data object and changes with each data object selection.



3. **LABELS PULLDOWN MENU**

Labels are generated from the data retrieved from the fields designated for label generation in the **<data set>.config** file. See the section, **Display Configuration**, for information about the configuration file.

The following selections for label generation allow the user to determine if the speed of label generation is more valuable than a better label representation of data clusters as the user navigates through the terrain. In other words, for a very large data set, a "**quick**" label generation may be more valuable to a user as he is zooming in on the data set. For a small data set, the user may be able to zoom into the area of interest faster when a "**good**" label generation selection is made which reflects a better representation of the underlying data.

Quick – use for very large data sets or if there are time constraints on the user during a VxInsight session

Medium – use for medium sized data sets for a better label generation and time is not a crucial factor

Good – best for smaller data sets or where there are no time constraints on the user

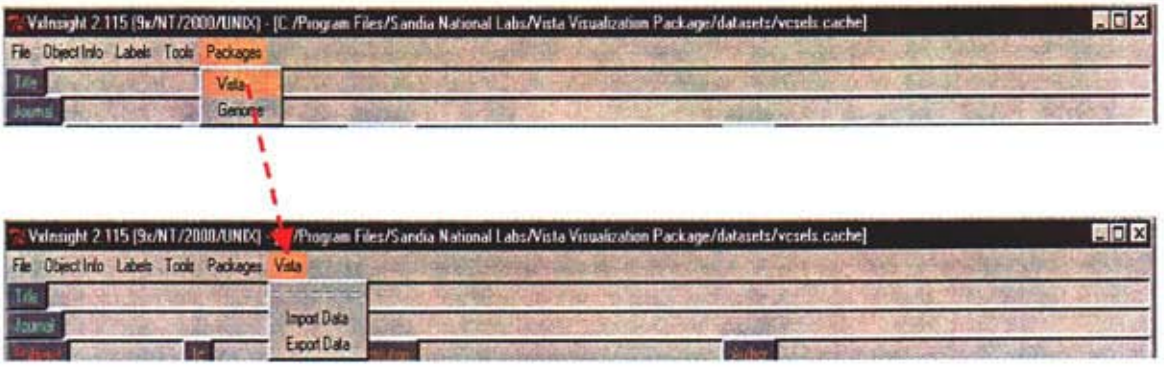
4. TOOLS PULLDOWN MENU

The "**Tools**" pulldown menu allows access to applications that are contained within the **VxInsight** environment. These are stand-along applications whose accessibility has been encapsulated within **VxInsight**.

5. PACKAGES PULLDOWN MENU

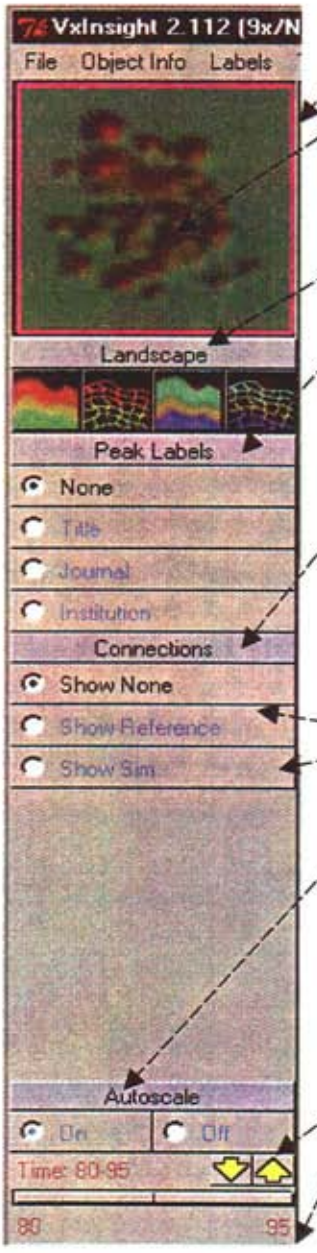
The "**Packages**" pulldown menu contains a set of customer-specific applications.

When selected, an item in the "**Packages**" list provides access these applications.



Customer specific packages will be included in the installation software with additional documentation with regard to their use.

B. SIDE MENU BAR



1. **Global View Window** – shows a user what portion (contained within the rectangle) of the entire landscape is currently being displayed. This feature is most beneficial when zooming and rotating the landscape to give the user an idea of the current area of exploration. A left mouse button click in the window expands it to include peak labels. Another click returns the global view window back to its original state.

2. **Landscape Icons** – used for visual display of data set

3. **Peak Labels** – a toggle menu used to determine which label field is used for peak labeling. The menu is generated using the label field names designated by a user in the **<data set>.config** file as discussed in the section, **Display Configuration**.

4. **Connections** - a toggle menu that allows the user to designate whether or not he wishes to show the relationships between specific data objects. The possible connections are determined prior to displaying a landscape from information obtained from the **<data set>.config** file as discussed in the section, **Display Configuration**.

   a. **Show None** – this is the default state when VxInsight is activated.

   b. **Example connection fields used**

6. **AutoScale** - used to keep the peak heights normalized

   a. **On** - landscape is continuously normalized height-wise to provide a uniform viewing of different scales of resolution.

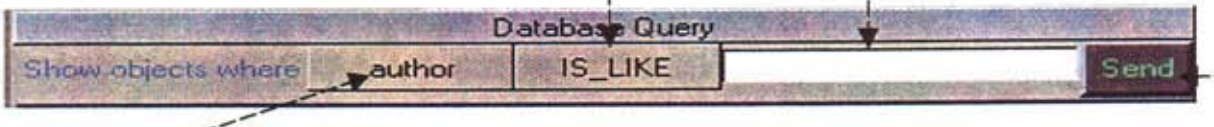
   b. **OFF** - used in conjunction with the range slider, as the slider moves over the range, the dramatic changes in the terrain as the data object density increases gives the user a quick view of major activity.

7. **Range** – allows a user to narrow the data set to a specific range of information. Upon start-up of a data set display, the range is determined if a range field has been defined in the **<data set>.db** file as discussed in the section **Display Configuration**.

C. BOTTOM MENU BAR

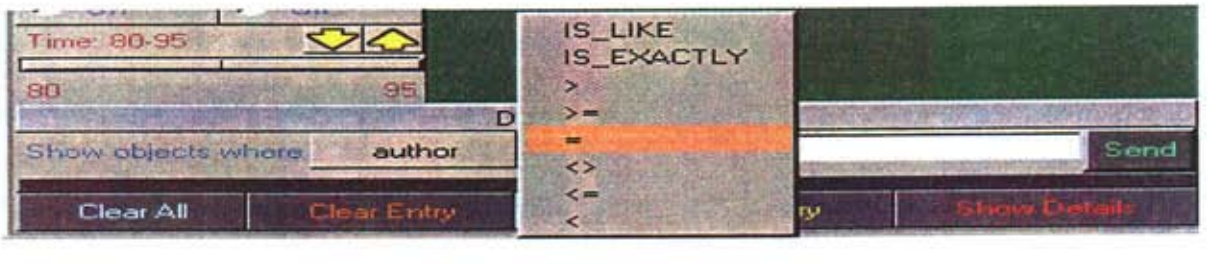
1. DATABASE QUERY

This menu allows a user to query the database for data objects containing specific information. These queries can be made on any of the fields specified in the **<data set>.db** file as described in the section, **Display Configuration**. See the section on **Database Information Retrieval** for instructions on its use.



Field – a specific field upon which the query depends. Once a data set is selected and displayed in VxInsight, this button will result in a "pulldown" menu listing all the possible fields from which to select.

Match Type – sets the match requirement when selecting a data object based on the input field value. The pulldown menu lists the choices that are available to the user.



Value – criteria value for selecting a data object. The input value can be as specific or as general as possible, but keep in mind that the more general the value, the more data the user will receive from the query.

Send – sends a query to the database requesting a list of all data objects matching the selection criteria

2. LEGEND

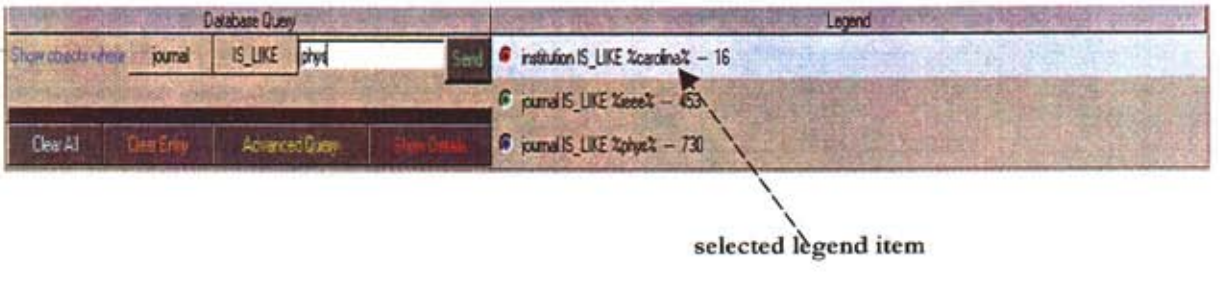
Displays specified criteria about each query made during a session. Each item in the legend is represented with a corresponding color that associates all the data objects on the landscape meeting the specified criteria. Specific data objects, shown as "pyramids", that have met criteria for more than one query will match the color of the latest query.

The following buttons control the legend display:

Clear All -- Clear all the legend entries. All data object color markers on the landscape, which are associated with each legend entry, will be cleared.

Clear Entry -- All data object color markers on the landscape, which are associated with the specific entry, will be cleared.

Show Details -- Displays a list of the contents of each data object meeting the specified criteria for a particular query. In order to show details on a query, select a query item from the legend and press the "**Show Details**" button.

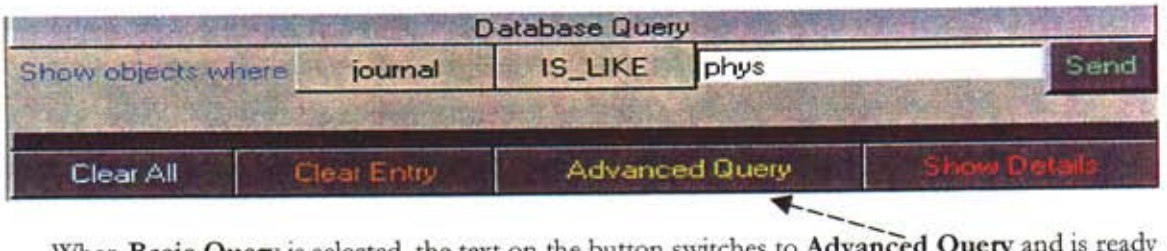


3. QUERY MODE

There are two types of database queries offered by VxInsight. One is a single or "basic" query and the other is a multiple or "advanced" query sometimes referred to as a "boolean" query.[3] A user can switch between "basic" and "advanced" mode using the query mode toggle button.

a. Basic Query Mode

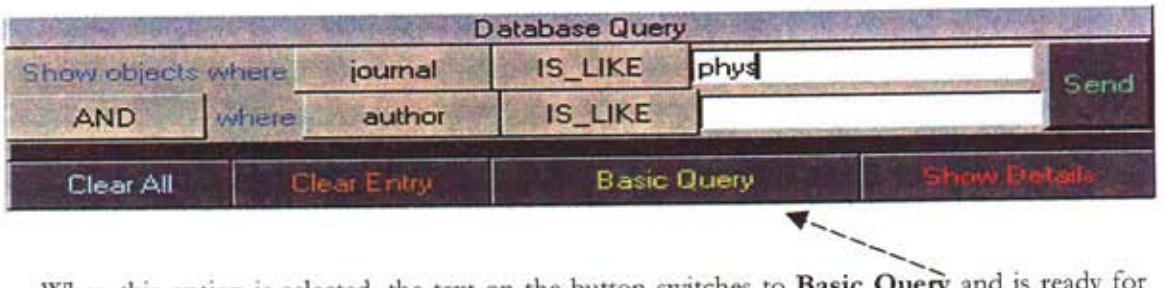
Allows a user to make a database query on a specific field of interest.



When **Basic Query** is selected, the text on the button switches to **Advanced Query** and is ready for selection by the user.

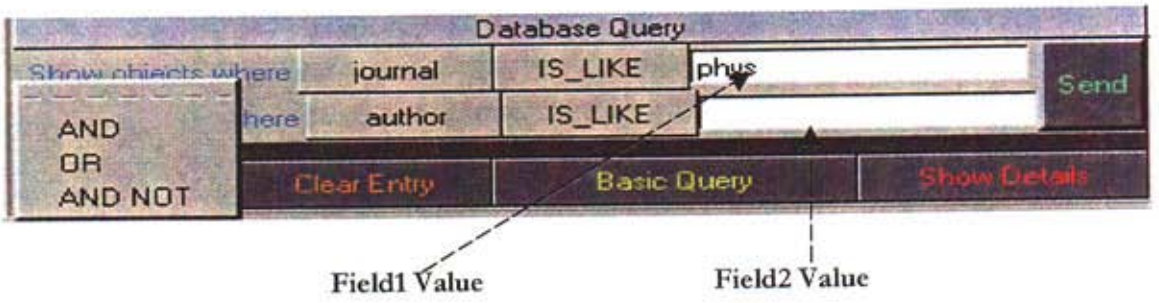
b. Advanced Query Mode

In **Advanced Query** mode, a user can make a query on two fields using a boolean operation to select data objects meeting the criteria.



When this option is selected, the text on the button switches to **Basic Query** and is ready for selection by the user.

The pulldown menu supports the following boolean operations:



[3] Only a two-field query is supported at this time.

AND – both values specified for the fields selected must be contained in a data object in order for it to be included in the resulting information

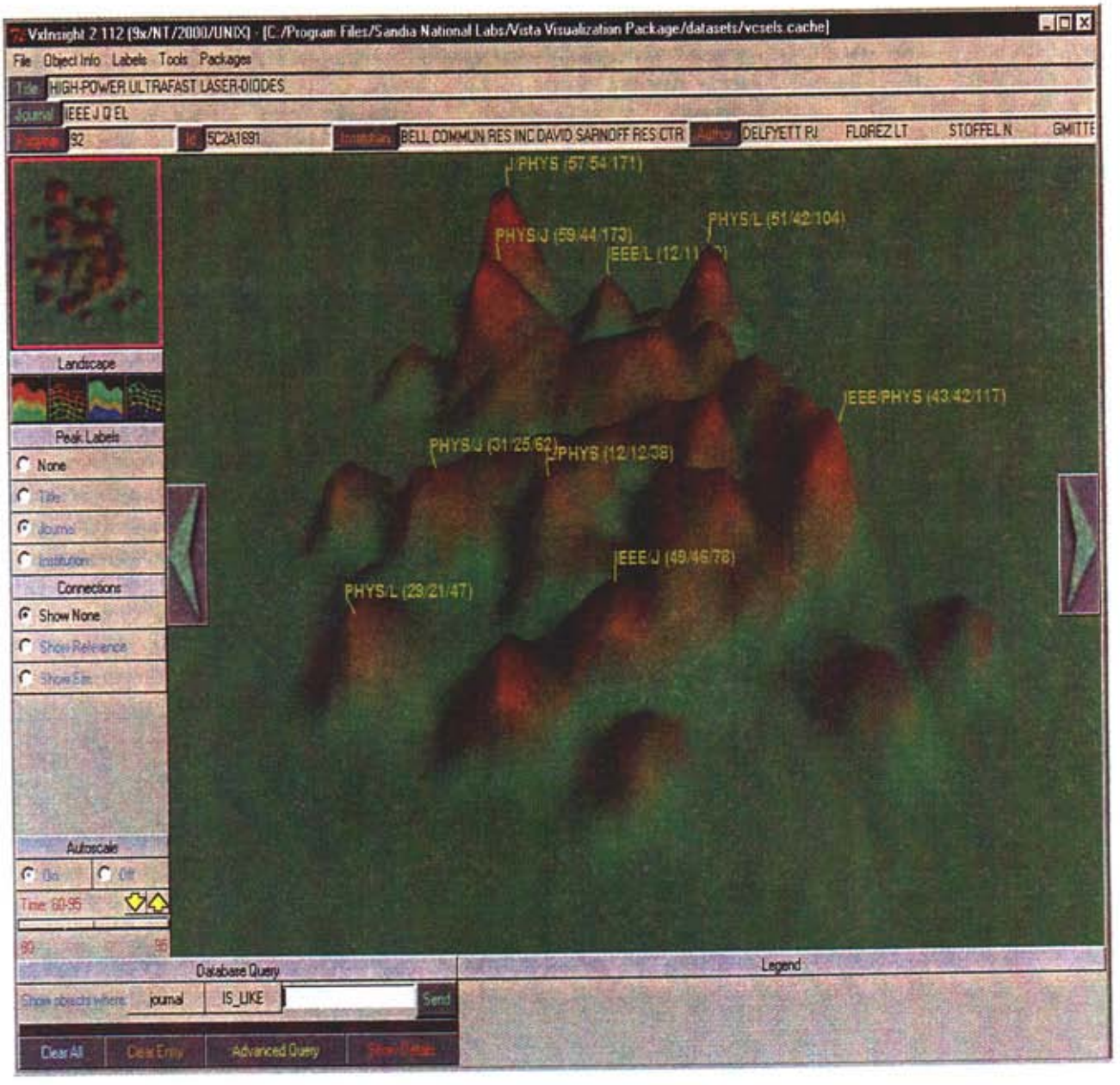
OR – either value specified for the fields selected must be contained in a data object in order for it to be included in the resulting information

AND NOT – the first but not the second value specified for the fields selected must be contained in a data object in order for it to be included in the resulting information

A. VISUAL DISPLAY

1. TERRAIN-VIEW MODE

The "mountain" peaks are generated based on the density of data objects in the plane. The height and conformation of a given peak is determined by the number of data objects contained within a specified radius on the plane.



Based on terrain-view mode, the field selection made from the **Peak Labels** menu, and the data objects contained within the radius of a specific mountain on the landscape, the peak labels are generated from the two most common values found in the field of all the data objects.

Label Format: <most common value>/<second most common value> (#/#/#)

where **(#/#/#)** is defined as

(number of occurrences of first value/number of occurrences of second value/total no. of data points sampled)

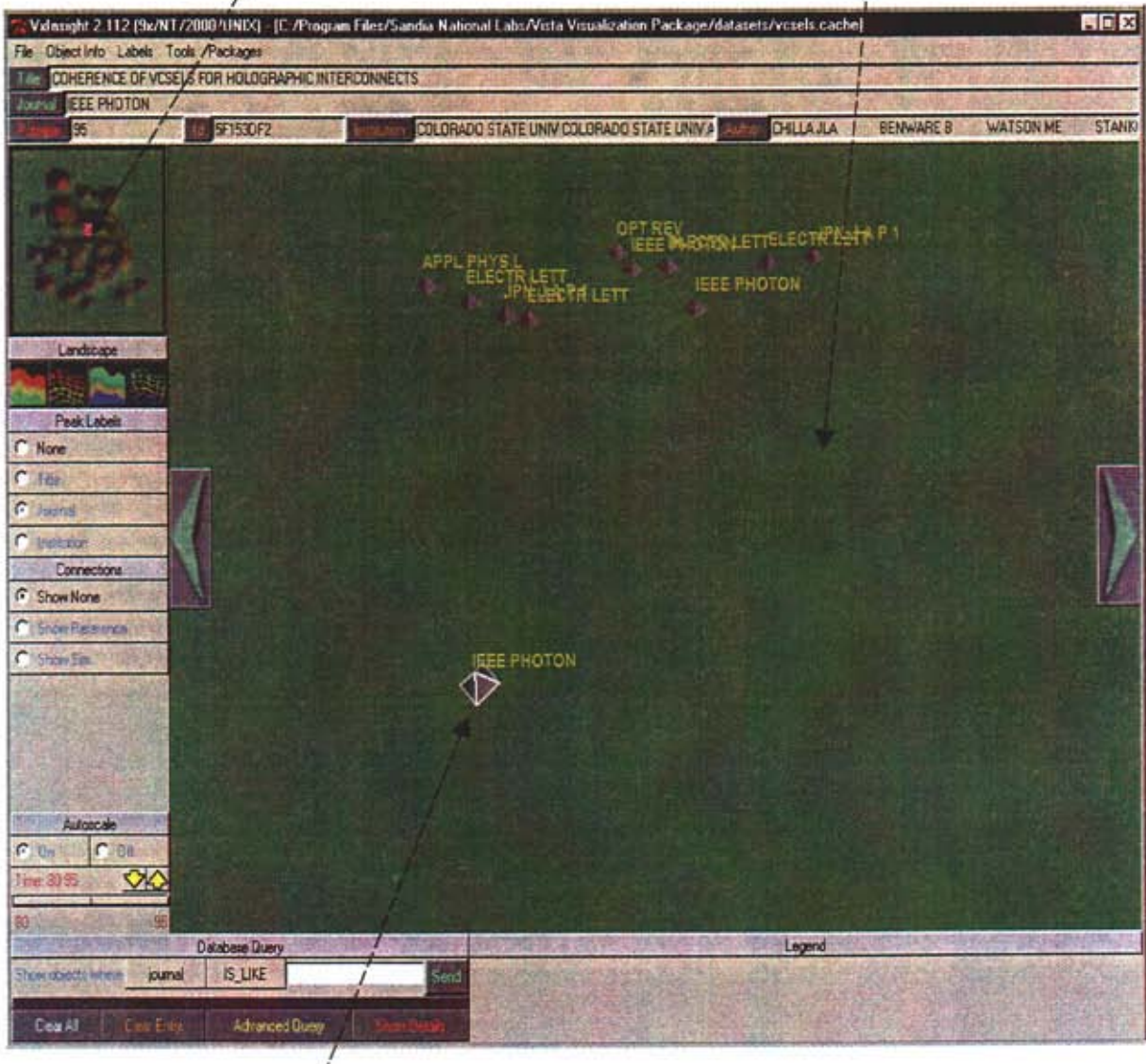
Example: Word1/Word2 (50/45/500)

Definition: Word1 occurred 50 times and Word2 occurred 45 times in a sampling of 500 data points contained within the radius of a specific mountain.

2. DATA RECORD MODE

As a user zooms down to "the bottom" of the mountain, the terrain changes until it reaches the individual data-objects display. At this point, each data object is displayed as a "pyramid" and represents a specific record of information.

The area within the rectangle in the global view window is represented in the landscape.



Based on data-object mode and the field selection made from the **Peak Labels** menu, each data object label is generated from the value found in the field for that data object.

As a user selects or "picks" a data object, the information menu fields are populated with values retrieved for the specific data object.

B. NAVIGATION

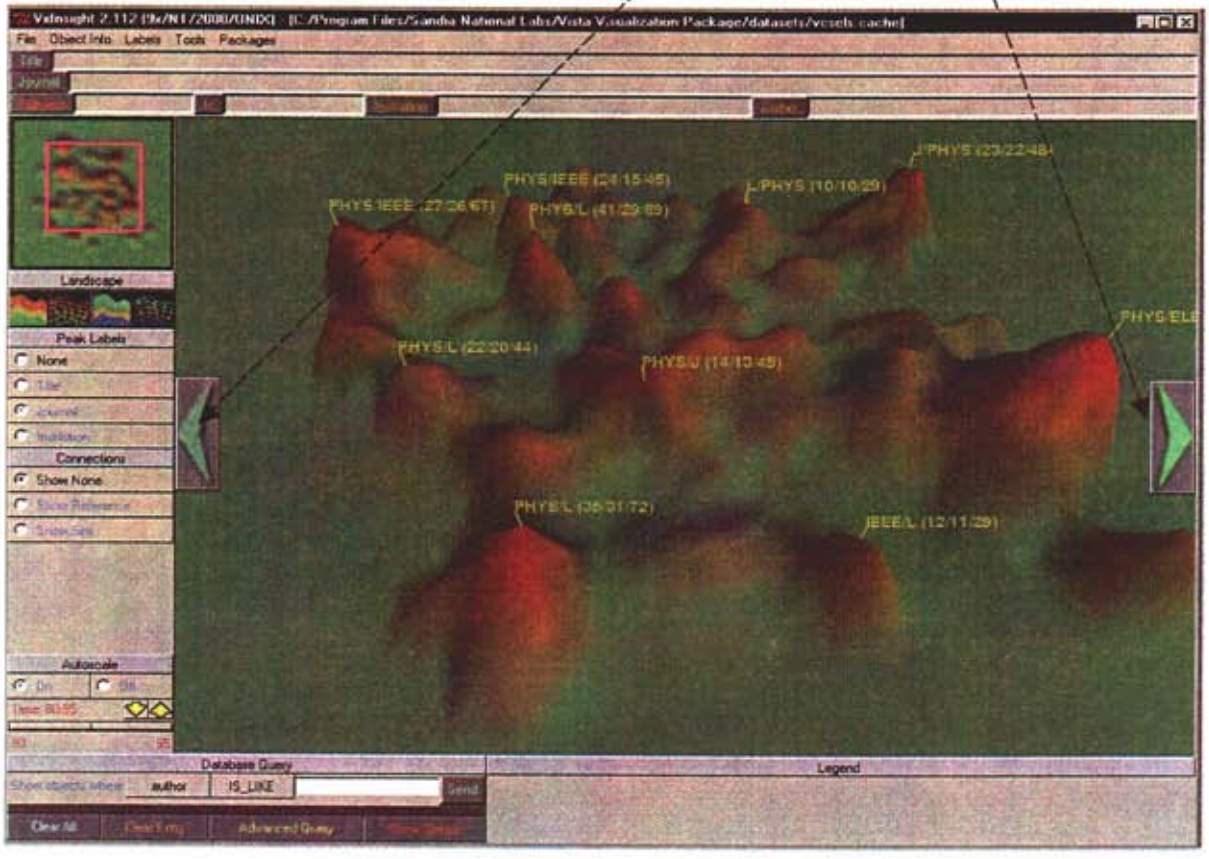
1. ZOOM

Zoom In – the left-mouse button click allows the user to zoom in to an area of interest. The user may click for a controlled zoom or may choose to hold the mouse button down for a faster zoom. As the user zooms in, the terrain label generation reflects the newly created peaks.

Zoom Out – the right-mouse button click allows the user to zoom back out. The user may click for a controlled zoom or may choose to hold the mouse button down for a faster zoom. As the user zooms back out, the dynamic label generation again reflects the newly created peak information.

2. ROTATE

There are two arrows supplied for rotating the landscape to the left or to the right. The user simply places the mouse pointer on an arrow and the landscape will begin to rotate in that direction. The highlighted arrow shown below signifies that the user is rotating the landscape to the right or clockwise.



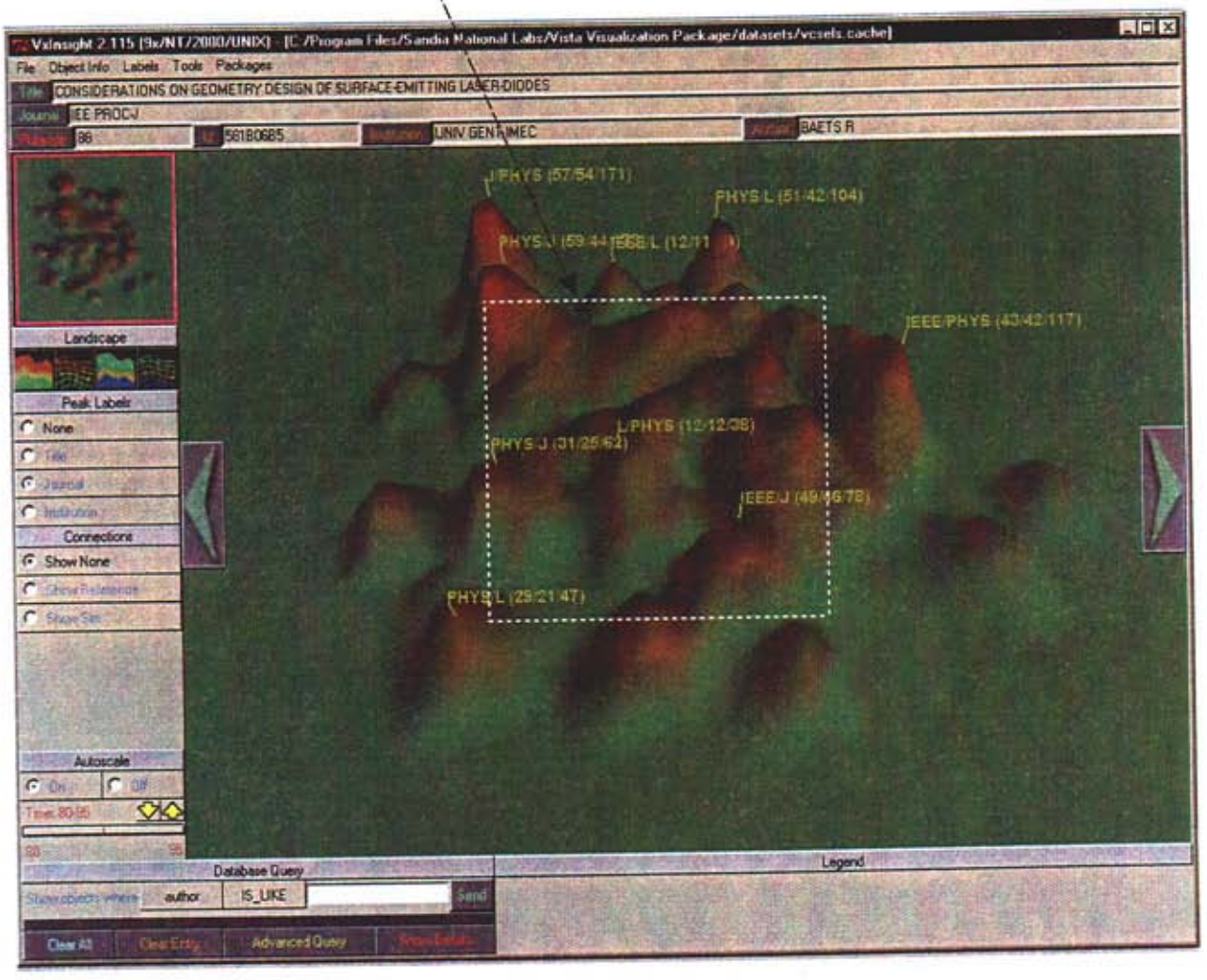
Specific database information can be retrieved through data object selection when in **"data record mode"** as described above or through database queries on a specified field or fields.

A. DATABASE QUERY

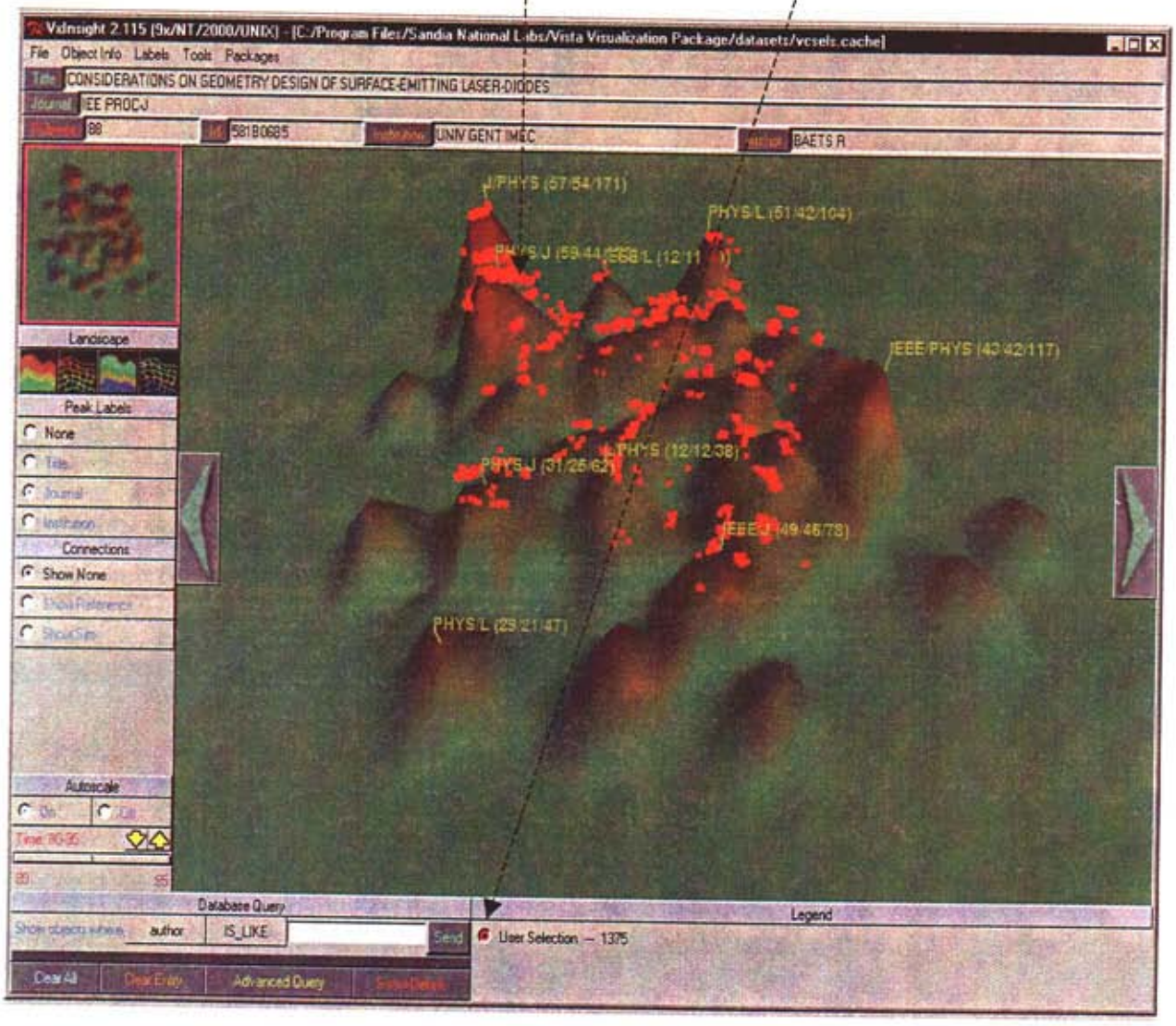
The **Database Query** menu described in the **MENUS** section, Item D1, is used to query the database for records containing specific information. These queries can be made on any of the table fields contained in the database file, **<data set>.db**, for the current session. The result obtained from the query is a set of **all** data object identifiers that meet the specified criteria.

B. SELECT DATA OBJECTS SUBSET

Another method of selecting a subset of data object information is by using the "bounding box" functionality. A user can encapsulate a portion of the landscape by using the "shift-click-and-drag" method that generates a bounding box around the area of interest on the landscape.



The selection database record references are marked and noted in the Legend menu.

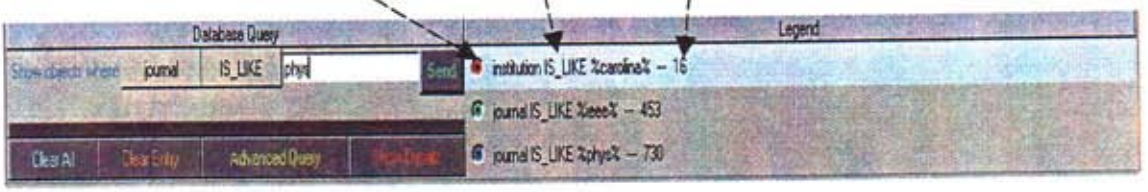


C. VISUAL DISPLAY OF QUERY RESULTS

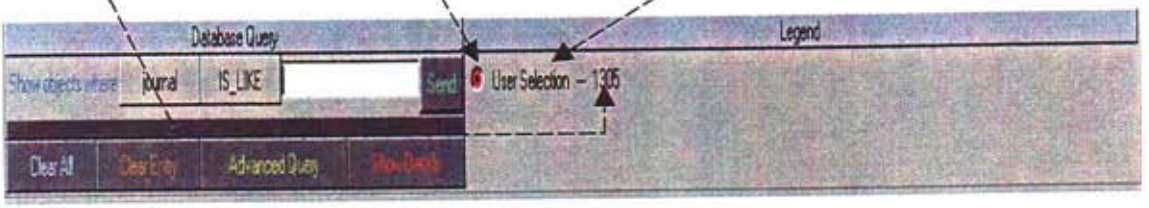
1. QUERY LEGEND

When query results are received through a database query or a bounding box selection, a unique color is assigned to the query in order to determine which objects matched which specified queries.

When database results are generated through a database query, the legend displays the unique color associated with the request text and the number of matching objects. The maximum number of entries on the legend is eight (8).

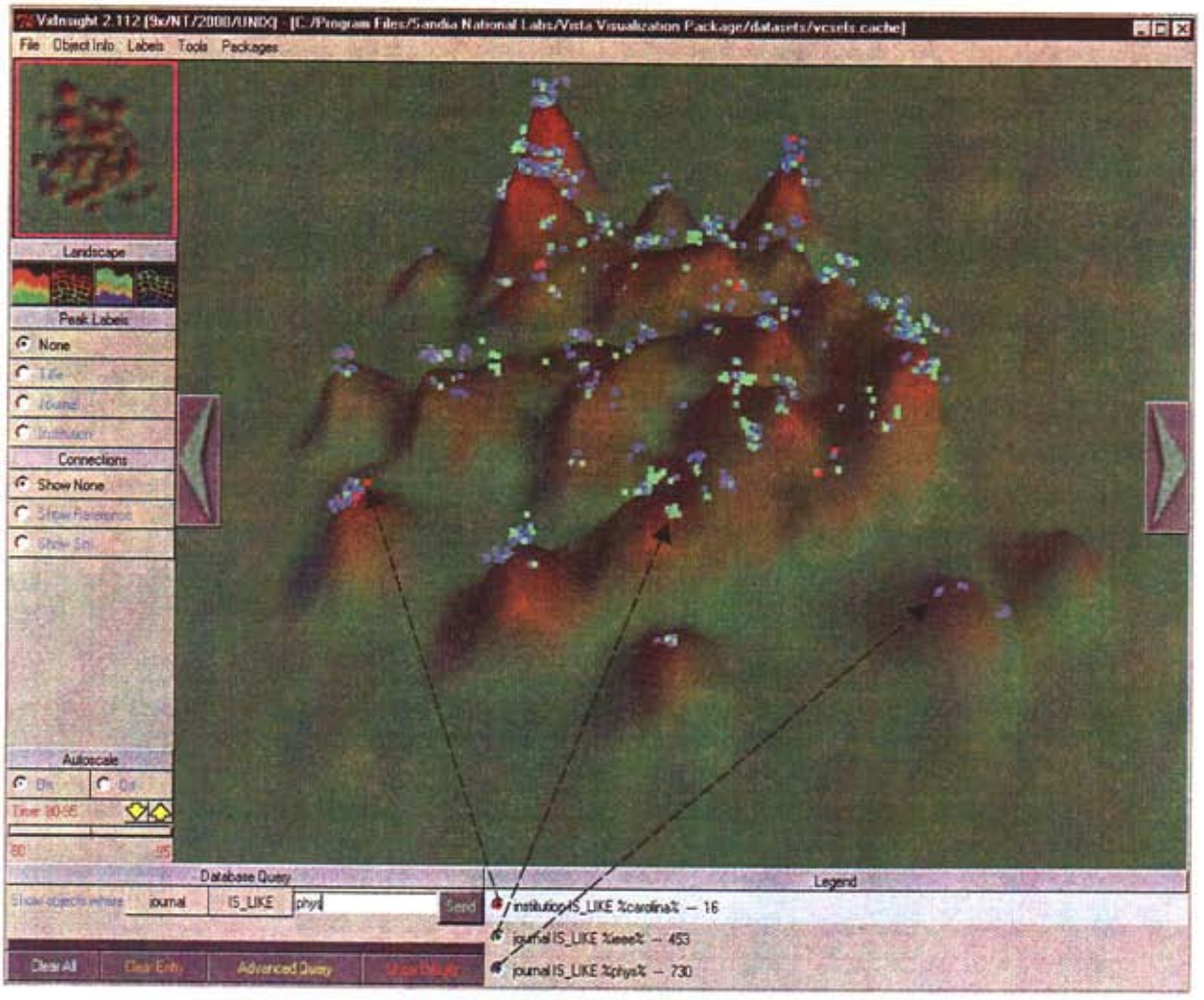


When database results are obtained through the "bounding box" selection process, the legend displays the unique color associated with the "User Selection" text label and the number of matching objects.



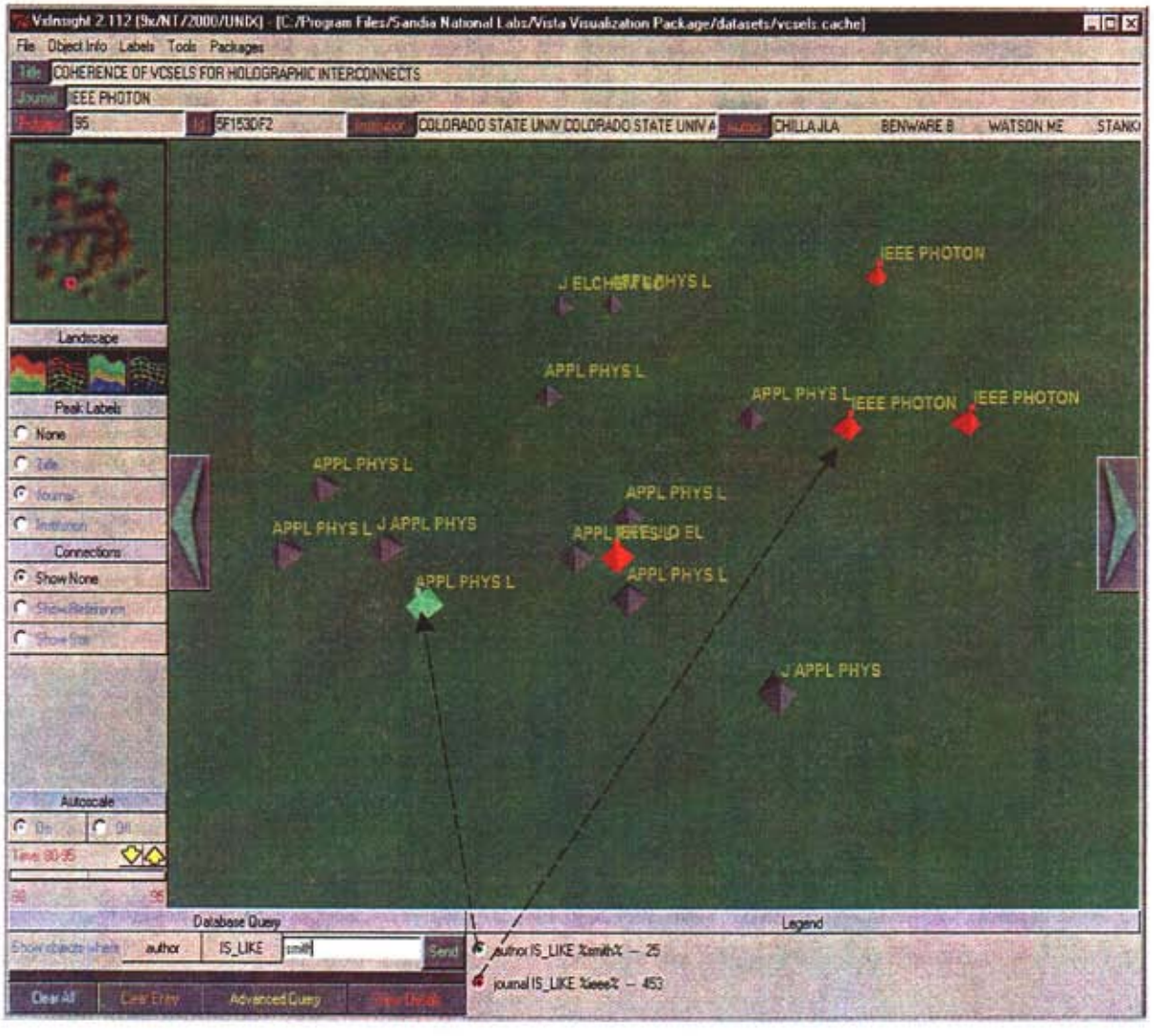
2. TERRAIN-VIEW MODE DISPLAY

The results of the database query or "bounding box" selection process are displayed as "data points" on the terrain matching the query "Legend" color. Each data point represents one data object meeting the specified query criteria.



3. DATA-RECORD MODE DISPLAY

In **data-record mode**, the color of the "pyramids" representing all the data objects meeting the specified criteria or contained within the "bounding box" is changed to match the color of the legend entry from which the data objects were selected.



GENERAL INFORMATION

This application was developed on a Windows NT platform using the Visual C++ programming environment. The graphical user interface was developed using Tcl/TK, version 8.0.

Questions should be addressed to:

Brian Wylie

Email: bnwylie@sandia.gov

Phone: 505-844-2238

DISTRUBUTION:

1	MS-0189	LDRD Office, 4001
1	MS-9018	Central Technical Files, 8945-1
2	MS-0899	Technical Library, 9616
1	MS-0612	Review and Approval Desk, 9612
1	MS-0318	K. W. Boyack, 9212
1	MS-0318	G. S. Davidson, 9212
1	MS-0318	P. D. Heerman, 9227
1	MS-0318	B. N. Wylie, 9227
1	MS-0321	W. J. Camp, 9200