Information Cartography Applied to the Semantics of Roget's Thesaurus

L. John Old

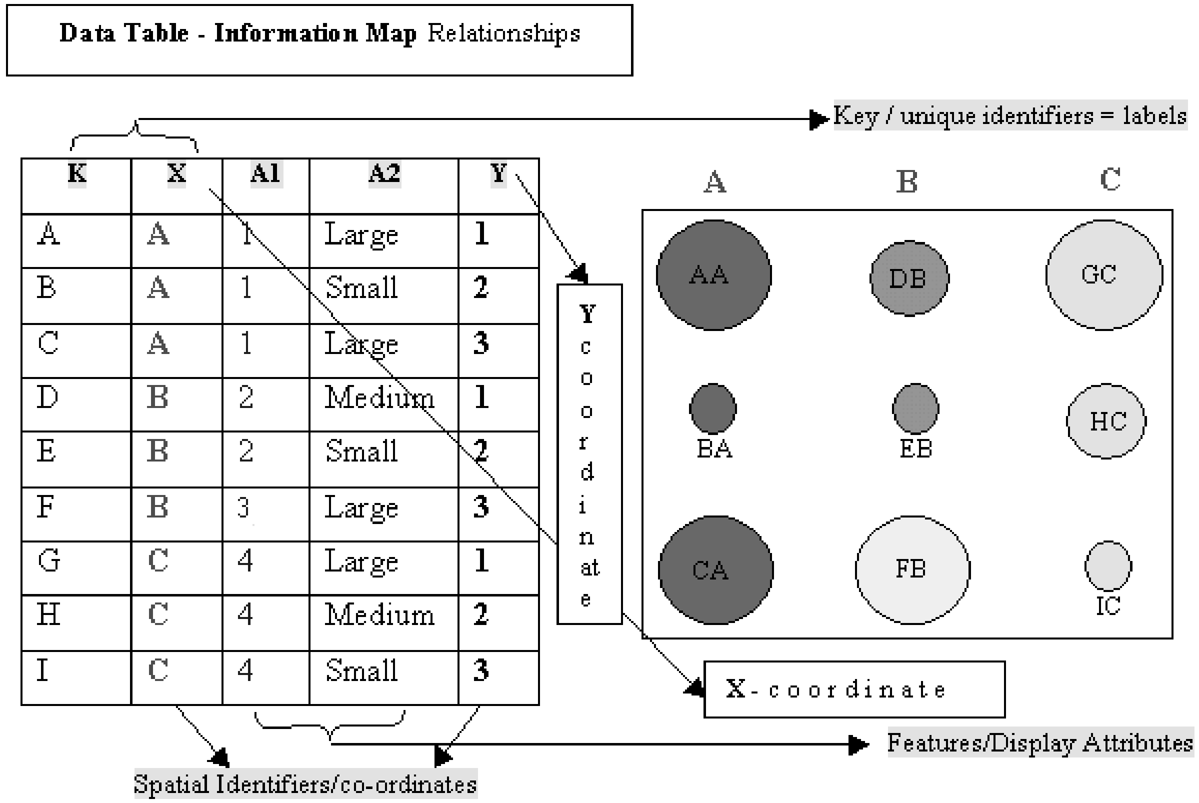
School of Library and Information Science
Indiana University
jold@indiana.edu

Abstract

Using algorithms, representation methods and models from Geographic Information Systems (GIS), information that has relationships between elements may be represented spatially, especially if some distance metric can be brought to bear. This paper discusses the use of spatial methods for the display of semantic data derived from Roget's Thesaurus, and shares some insights derived by using these methods.

Introduction

Visualization of traditionally non-visual information (statistics, lists, tables and the like) is a growing art in such fields as data mining, geo-spatial information processing, chemistry and pharmaceuticals, medical diagnosis, and economics. Visualization usually relies heavily on the spatial representation of non-spatial data. The main problem for non-spatial-data visualization is how to identify or extract existing attributes in the data that can be used for spatial representation, or, alternatively, how to convert the data to a form that has spatial attributes. The use of distance metrics (dissimilarity, relevance, disutility, and so on) which are then converted to a spatial format using multidimensional scaling (MDS) techniques, a common method. Other possibly fruitful methods of conversion are factor analysis and Kohonen nets (Small, 1998), clustering and geometric triangulation (Small, 1999), and singular-value decomposition. The examples discussed in this paper use a scatterplot method.


Data Table - Information Map Relationships
Key / unique identifiers = labels
A
B
C
AA
DB
GC
BA
EB
HC
CA
FB
IC
1
2
3
Y coordinate
X- coordinate
Features/Display Attributes
Spatial Identifiers/co-ordinates

K	X	A1	A2	Y
A	A	1	Large	1
B	A	1	Small	2
C	A	1	Large	3
D	B	2	Medium	1
E	B	2	Small	2
F	B	3	Large	3
G	C	4	Large	1
H	C	4	Medium	2
I	C	4	Small	3

Figure 1 Data Table – Information Map relationships

This paper aims to demonstrate the power of spatial information systems to represent semantic data in new and useful ways. Westerman suggests that this may be a natural way to represent semantic information.

> …diverse, non-spatial information can be represented within a spatial context. Some of the mechanics that underlie this mapping have been proposed by

> Jackendoff (1983; see also Gardenfors, 2000) who argues that the highly developed capacity of the human brain for spatial processing is responsible for the application (during the developmental process) of similar structures to the cognitive organization of information from other semantic fields. Consequently, the semantic primitives that describe spatial associations (motion and location) are held to form a superset from which associations in any other semantic field can be described. (Westerman, 2000, [Cognitive Processes]).

Westerman concludes that the implication of this position for the development and use of visual information systems is that any given semantic dimension of computer-stored information can be represented in a spatial format, and that any computerized information space can be navigated using similar cognitive processes to those that would apply during the process of 'real world' navigation.

Geographic information systems (GIS) can be used to both store data and information (data that has been interpreted, or has meaning) and visualize it as 'information maps.' GIS facilitate dynamic display of, and interaction with, the data by creating relationships between the data and display elements. Figure 1 demonstrates the relationship between the data model and the visual structures in an information map. The relationship is simpler (but perhaps less flexible) than for general information visualization as the data objects have fixed co-ordinates derived from the data table (X, Y attributes, in this table). The identifier for entities is usually a label, in this case the compound key index, K+X. In a geographic map these would be labels for landmarks, such as cities or states. Any of the information in the table may be used as display variables. In this case A1, a numeric attribute, has been represented by colors; and A2, an ordinal scale that refers to the entity, is used to map entity size to circle size. Statistical packages such as SPSS can be used to create the X, Y coordinates via the multi-dimensional scaling (MDS) option. GIS systems such as ESRI's ArcView can be used to import and manipulate the

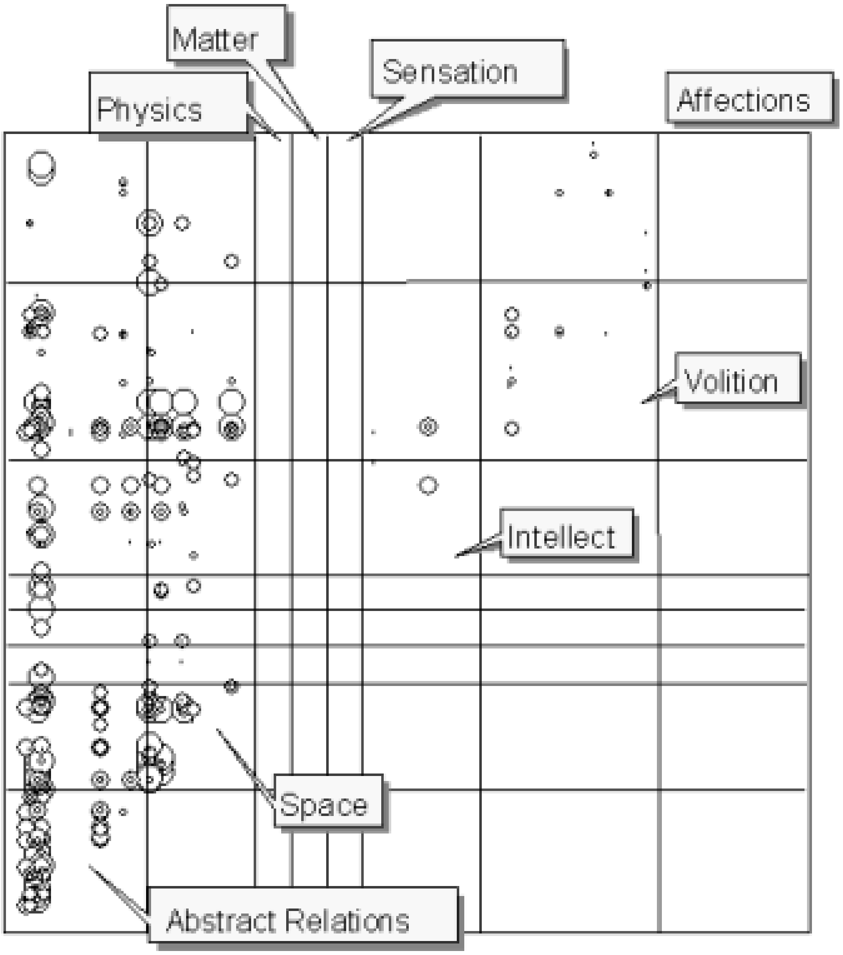


Figure 2. Macro-structure of Roget's Thesaurus

data as maps. ArcView was used to create the graphics presented here, but Open Visualization Data Explorer (OpenDX) (an "open source" no-charge-for-use visualization system (IBM, 1999)) is one of many alternatives.

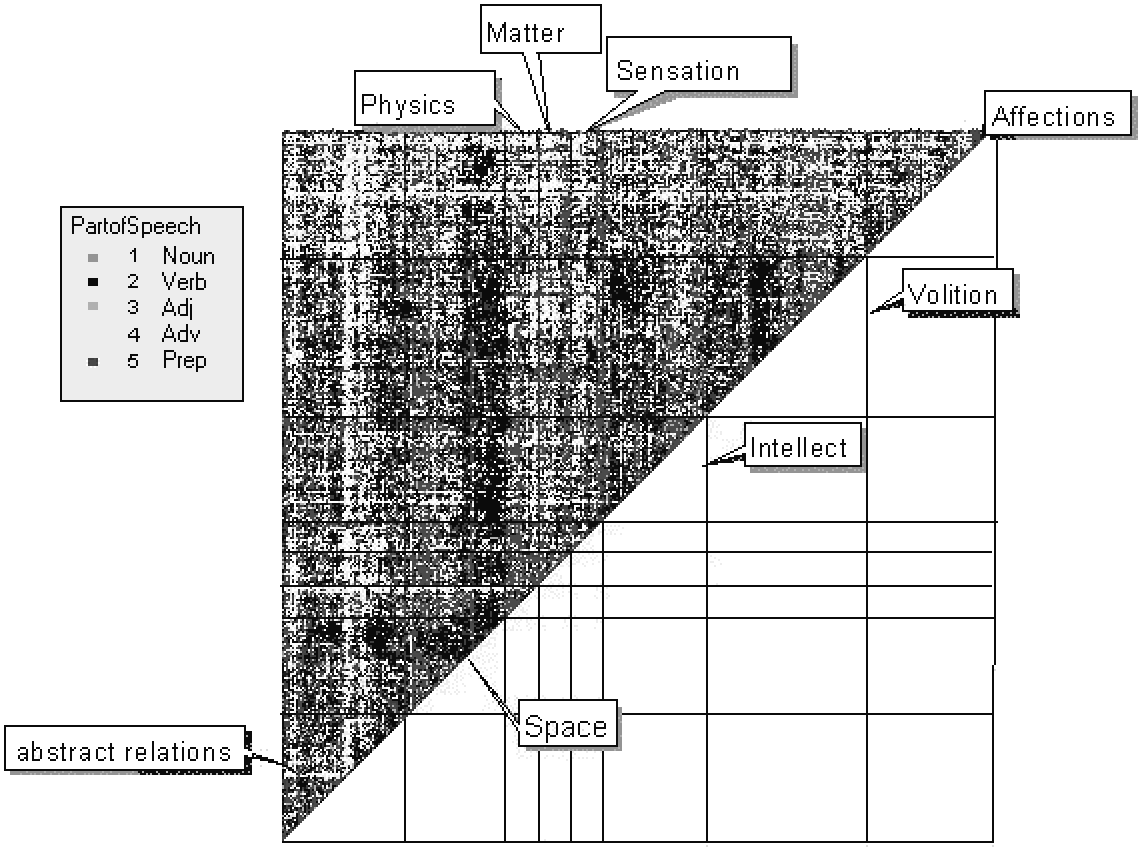


Figure 3. Part-of-speech distribution in Roget's Thesaurus.

Discussion

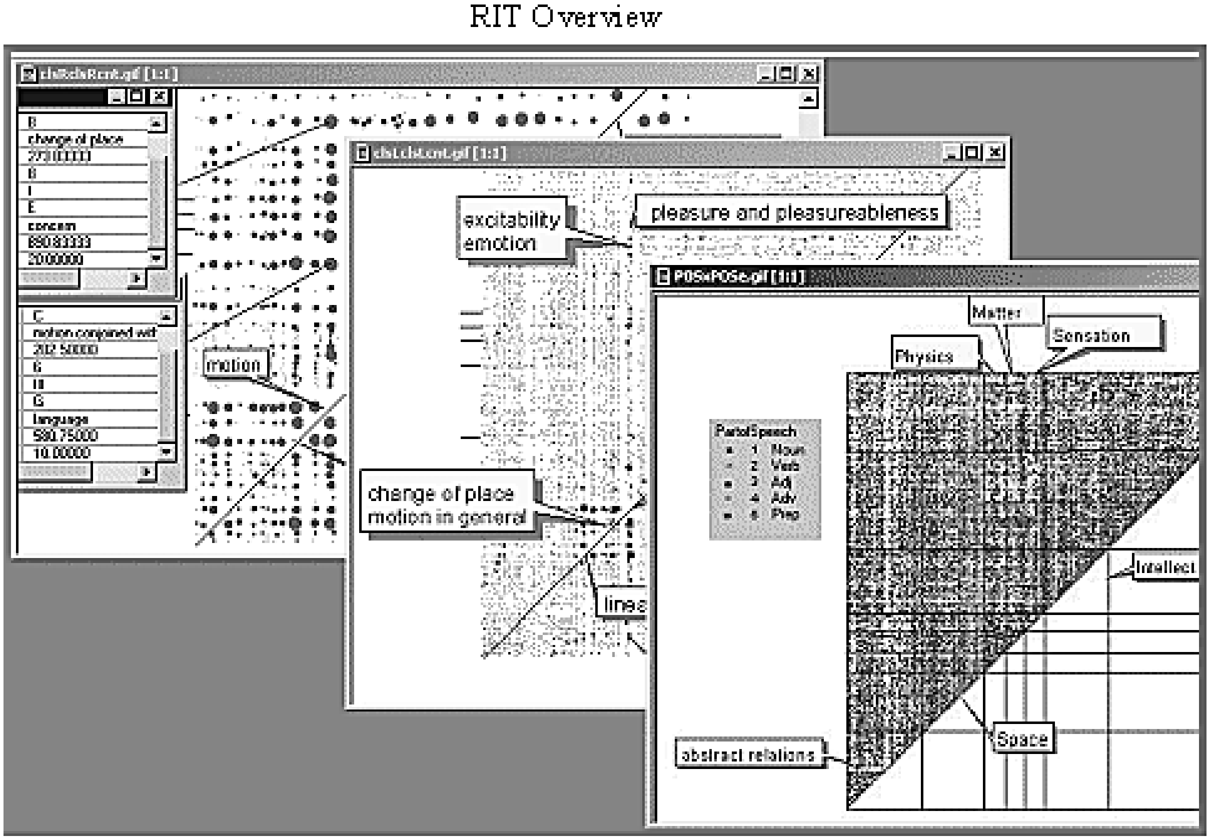
The process of developing information visualizations using GIS technology, cartographic principles and spatial metaphors is called "information cartography" (Old, 2001). The following examples demonstrate information cartography applied to data from a machine-readable version of Roget's Thesaurus.

Figure 2 shows the macro-structure of Roget's International Thesaurus (RIT) using the Roget top level classes (RIT is a conceptual hierarchy) as both X and Y coordinates. Sub-classes and categories down to the level of synonym groups (or synsets, in WordNet terminology) can be nested within the classes using relative coordinates. The points shown on the map in Figure 2 represent the locations of prepositions in RIT (concentrated in the 'Abstract Relations' and 'Space' classes). The difference in size among the points represents a polysemy scale. GIS also allow the display of bar graphs (or pie charts) of data in place of points or symbols.

Figure 3 shows the relationship between synsets, as reflected by shared words. The points, representing words, are color-coded by part-of-speech, and it can be seen, even in black-and-white, that a band of verbs dominate the relationship between the 'Space' and 'Volition' classes.

Figure 4 shows the relationship between areas of RIT at three different levels of the hierarchy. These can be viewed as overlapping maps in a GIS. The left (rear) map is labeled with the sub-class of *Space*, 'Motion.' The middle map (at a lower level of abstraction in the Roget hierarchy) is labeled with two sub-classes of *Motion*, 'Change of Place' and 'Motion in General,' identifying the areas where the most polysemous words exist (a close-up of the verb-concentration identified in Figure 3). The small windows in the rear map demonstrate how data associated with any point may be retrieved on demand.

Based on connectivity analysis of RIT and correlations with Indo-European roots, Old (2000) has proposed that language originated from alarm calls of human ancestors. The fright-fight-flight-freeze dimensions of the alarm response can be seen in the multi-dimensional scaling of Type-10 chains (Bryan, 1973), the strongest associations in RIT.



B
change of place
77.000000
B
I
E
concern
890.83333
20.00000

C
motion conjoined wit
202.50000
6
III
G
language
590.75000
10.00000

Figure 4. Comparing hierarchical levels as layers

According to George Miller (Miller et al., 1993) it has long been known that frequency of occurrence and polysemy are correlated. That is to say, on the average, the more frequently a word is used the more different meanings it will have. Furthermore, common sense dictates that the longer a word is around, the more senses it is likely to accumulate. It is also likely that associations between words--relations such as synonymy--also increase over time. Consistent with this is Joseph Novak's observation that "Meaningful learning involves the assimilation of new concepts and propositions into existing cognitive structures" (Ausubel, Novak, & Hanesian, 1978).

To the extent that words represent concepts, this suggest that our most polysemous and connected words reflect concepts that have been around the longest. Figures 3 and 4 support the fight-fight theory in that they identify the largest clusters of polysemous words in RIT as being verbs related to change of place.

This work on visualization builds on previous work on the visualization of Roget's International Thesaurus (Old, 1999a), which focused on representing word fields of individual words (semantic neighborhoods); and of citations between authors (Old, 1999b), which focused on influence or affiliation of individual authors. Figure 5 shows the semantic field of the word "over" as a topographical landscape. The large "hill" in the center represents *over* and the surrounding points represent words that occur as synonyms of over in RIT, arranged using a distance metric (Old, 1996) and multi-dimensional scaling. The topology is derived by a GIS which first generates elevation contours derived from values associated with each point (in this case, number of senses shared with *over*), then fills the gaps between contours with a triangular irregular network (TIN) that appears as a surface.

Similar information to that in Figure 5 can be derived from the representation described in Figures 2 through 4 by zooming in to small areas, or selecting entries (particular senses of a word) using an SQL query of the data table associated with the map. The results are highlighted by the GIS both in the map and in the source table, and can be exported or added back in as separate map layers.

Information can alternatively be derived by selecting features using a spatial query. A spatial query in its simplest form consists of selecting features by dragging the mouse. Utilities built into GIS allow also for the spatial selection of features through the intersection or union of features within different layers of a map.

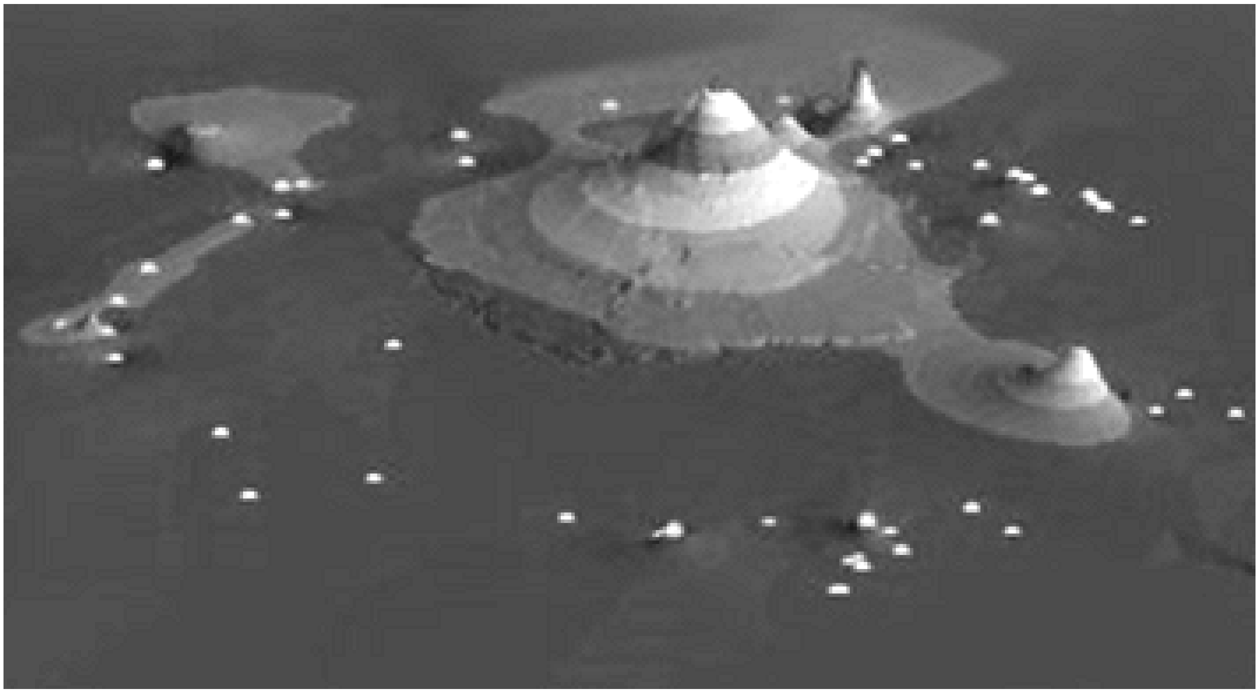


Figure 5. Topographical representation of a semantic field.

GIS can manipulate other data-types--for example polygons, which represent data as areas; and lines, which represent relationships between points as in graphs. Even automatic categorization methods (not as yet fully evaluated) may be applied by using GIS buffering utilities.

Conclusion

This paper has introduced a cross section of methods available for non-spatial-data analysis in the spatial information systems domain. The data analysis method introduced here demonstrates just some of the many available options; GIS allow for the manipulation of three-dimensional maps enabling rotation and inspection from different angles. Even the two-dimensional maps allow zooming, panning, and the hiding of data layers during inspection. Three-dimensional maps can also be exported as VRML files, viewable in virtual reality environments or web browsers.

These methods are applicable to a wide range of non-spatial data but some limitations still exist. Apart from the use of animation, time series data are difficult to deal with in a GIS. The development of a conceptual model and associated tools for the visualization of spatial-temporal process information is among the goals of the Commission on Visualization of the International Cartographic Association (ICA, 1997).

The data in spatial information systems are most easily interpreted in color graphics; interpretation of the figures here in gray-scale graphics, is difficult. Another disadvantage of these black and white pictures, as opposed to color pictures, is that less information can be displayed. Lines, analogous to roads or rivers in a map, may be used to define relationships between entities and allows for the use of lattice and graph methodologies (Priss and Old, 1998).

Further analysis and evaluation of these methods will be conducted in terms of usability, the evaluation dimensions for Visual Information Retrieval Interface (VIRI) (Rorvig & Hemmje, 1999), and Tufte's (1983) principles of graphical excellence.

References

Ausubel, D. P., Novak, J. D., and Hanesian, H. (1978). *Educational Psychology: A Cognitive View,* 2nd Edition, New York: Holt, Rinehart and Winston. Reprinted, New York: Warbel and Peck, 1986.

Bryan, R., (1973). Abstract Thesauri And Graph Theory Applications To Thesaurus Research. In Sally Yeates Sedelow, editor, *Automated Language Analysis*. University of Kansas Press.

ESRI (1999). *ArcView GIS*. Available: http://www.esri.com

International Cartographic Association Commission on Visualization (August 1997), *Overview*. Available: http://www.geovista.psu.edu/icavis/

IBM (1999). *Open Visualization Data Explorer*. Available: http://www.research.ibm.com/dx/

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. (1993). *Introduction to WordNet: An Online Lexical Database (Revised August 1993)*. Available: http://www.cogsci.princeton.edu/~wn/papers/

Old, L. John, (1996). Synonymy and Word Equivalence. *Online Proceedings of the 1996 Midwest Artificial Intelligence and Cognitive Science Society Conference* (MAICS96), Bloomington, IN.

Old, L. John, (1999a). *Spatial Representation of Semantic Information.*MAICS99 presentation notes. Available: http://php.indiana.edu/~jold/maics/maics.htm

Old, L. John, (1999b). *Spatial Representation and Analysis of Co-Citation Data on the "Canonical 75": Reviewing White and McCain.* Available: http://php.indiana.edu/~jold/SLIS/L710.htm

Old, L. John, (2000). Core Concept Patterns in English Semantic Networks and Indo-European Roots. Proceedings, *Connections 2000: The Sixth Great Lakes Information Science Conference*. Knoxville, TN.

Old, L. J., (2001). Utilizing Spatial Information Systems for Non-Spatial-Data Analysis. *Scientometrics*, Vol. 51, No. 3 (2001) 563–571.

Priss, U., and Old, L. J. (1998). Information Access through Conceptual Structures and GIS. In *Information Access in the Global Information Economy. Proceedings of the 61st Annual Meeting of ASIS*, 1998, pp. 91-99

Roget's International Thesaurus, 3rd Edition, Thomas Crowel Company, 1963.

Rorvig, M., and Hemmje, M. (1999). Conference Notes--1996: Foundations of Advanced Information Visualization for Visual Information (Retrieval) Systems. *Journal of the American Society for Information Science*. 50(9): 835-837.

Small, H., (1998). Personal communication.

Small, H. (1999). Visualizing Science by Citation Mapping. *Journal of the American Society for Information Science*. 50(9): 799-813.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT.: Graphics Press.

Westerman, S. J., (2000). *Information Retrieval / Visualisation and Related Publications: Virtual Information Spaces*. Available: http://www.human-factors.org.uk/inf_ret.htm

Westerman S.J., Cribbin T. (2000). Mapping semantic information in virtual space: Dimensions, variance, and individual differences. *International Journal of Human-Computer Studies*, 53, 631-866.