Colloquium

The world of geography: Visualizing a knowledge domain with cartographic means

André Skupin*

Department of Geography, University of New Orleans, New Orleans, LA 70148

From an informed critique of existing methods to the development of original tools, cartographic engagement can provide a unique perspective on knowledge domain visualization. Along with a discussion of some principles underlying a cartographically informed visualization methodology, results of experiments involving several thousand conference abstracts will be sketched and their plausibility reflected on.

The question "Hasn't everything been mapped already?" is commonly posed to someone who calls himself a cartographer in the early 21st century. It would then typically be countered with reference to the ever-changing nature of what geographers like to call the "infinitely complex geographic reality," requiring vigilance in keeping ever-more-detailed geographic databases up-to-date. Where ever-growing geospatial data repositories, advanced computing power, and cognitive insights meet, cartographers are advancing scientifically in a field known as geographic visualization.

At the fringes of this activity, some cartographers have begun to attempt a combination of centuries of accumulated cartographic knowledge with modern computational approaches and cognitive insights, toward the visualization of nongeographic information. Examples for such nongeoreferenced data are the text document corpi held in digital libraries, user interaction logs created by Web applications, or biological data associated with genome mapping. In all of these cases, researchers of the interdisciplinary effort known as information visualization are engaged in the endeavor of making high-dimensional structures more directly accessible to the human cognitive system (1). Arguably, lessons from traditional cartography and transformation techniques derived from geographic information science would be applicable to many aspects of information visualization (2). This holds especially true in the context of 2D representations on screen or paper and in the even more narrowly defined, yet extremely popular, group of map-like information visualizations (3). Some results of this ongoing cartographic involvement are discussed here.

Implementation of Map-Like Knowledge Domain Visualization

A spatialization of the geographic knowledge domain is presented here on the basis of an analysis of abstracts submitted to the annual meeting of the Association of American Geographers. With all of the branches of geography participating at this meeting, the data set and resulting visualizations provide a fairly comprehensive snapshot of the geographic discipline, from established, well-publicized research fields to those only recently emerging. The goal is to implement a multilevel visualization, in which major research areas as well as finer nuances of geographic activity would be shown. There is a range of possible uses for such visualizations. Beginning geography students could be introduced to the topical structure of the discipline. Geographic researchers could see their own work in the context of broader disciplinary trends. Visualizations like this could ease collaboration in interdisciplinary research settings, and so forth.

The input data set consisted of 2,220 abstracts, as submitted by conference participants, stored in ADOBE pdf format on the conference compact disk. After conversion to a plain text format, each abstract's content was parsed into such components as title, author information, full text, and author-chosen keywords. Then, the text of abstracts was indexed against the full set of author-chosen keywords of all abstracts (4).

In the absence of citation information, the visualization methodology chosen in this project follows a straightforward content-based path (as opposed to exploiting an explicit citation link structure) based on vector-space modeling and use of the self-organizing map (SOM) method (Fig. 1). The methodology is thus related to a number of projects using a similar approach (5–7). However, there are also significant differences that, in combination, lead to visualizations bearing a distinctly cartographic mark (Fig. 2).

Following a standard vector-space implementation for the document corpus, a SOM consisting of a relatively large number of neurons is trained (4,800 neurons) so that unique 2D locations for individual documents can be derived (2,220 documents). Then, a hierarchical cluster solution involving all n-dimensional neuron vectors ($n = 741$) is computed to support a multiscale zoomable visualization (4).

Because natural language is the primary means by which scientific knowledge is formally disseminated and conveyed in many domains, meaningful labeling of geometric features ought to be not an afterthought but an integral part of knowledge domain visualizations. Contrary to common performance-oriented level-of-detail approaches, the aim here is to convey semantic aspects of the geographic domain in accordance with scale-dependent notions of global vs. regional vs. local structures. For example, the distinction of human geography and physical geography is a global one, whereas urban and industrial geography are regional flavors of human geography, and abstracts dealing with car manufacturing locations across the globe would form local structures. To this end, the extraction of scale-dependent label terms is particularly stressed. Determination of cluster labels is based on a weighting formula that extends the popular term frequency × inverse document frequency mechanism from its traditional use for individual documents (8) toward groups of documents. For a given cluster, this formula will tend to emphasize terms that appear often within and rarely outside of that cluster, accommodating very well the needs of a multiscale representation. When dealing with a small number of clusters (i.e., at a global scale), the derived label terms will be quite general, e.g., "climate" or "urban." For a large number of

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviation: SOM, self-organizing map.

*E-mail: askupin@uno.edu.

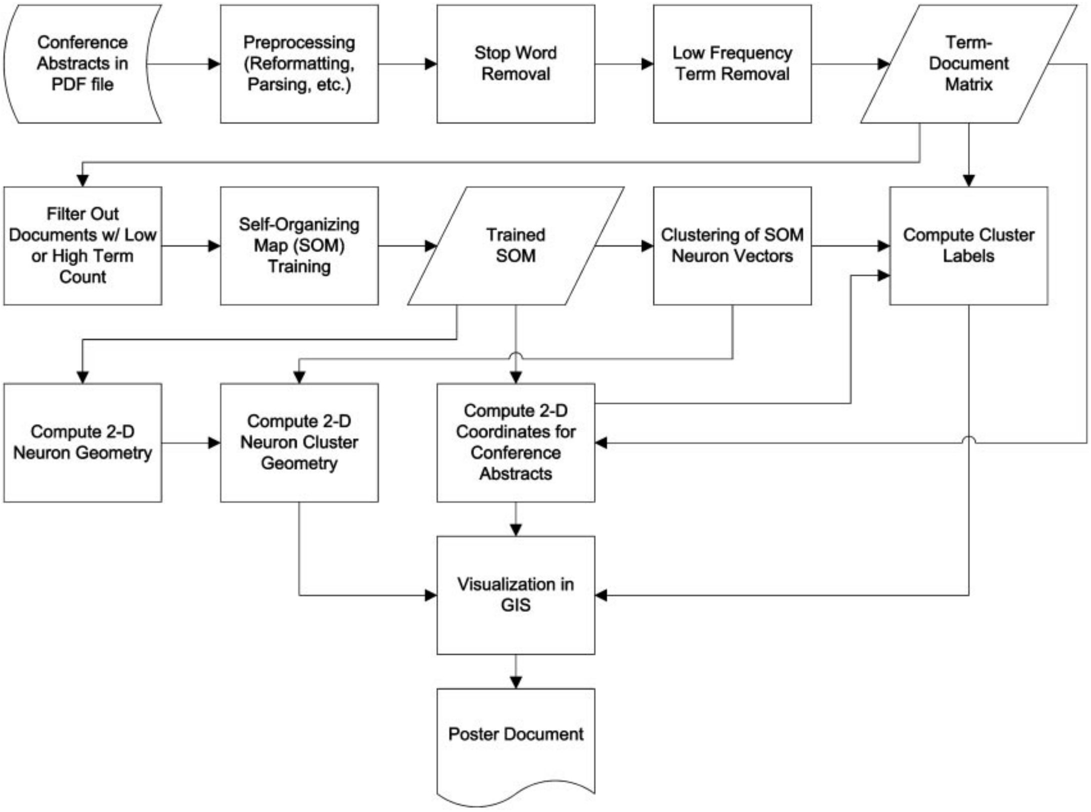
Conference Abstracts in PDF file
Preprocessing (Reformatting, Parsing, etc.)
Stop Word Removal
Low Frequency Term Removal
Term-Document Matrix
Filter Out Documents w/ Low or High Term Count
Self-Organizing Map (SOM) Training
Trained SOM
Clustering of SOM Neuron Vectors
Compute Cluster Labels
Compute 2-D Neuron Geometry
Compute 2-D Neuron Cluster Geometry
Compute 2-D Coordinates for Conference Abstracts
Visualization in GIS
Poster Document

Fig. 1. Creation of a map-like visualization of conference abstracts using a self-organizing map and geographic information systems.

clusters (i.e., at a local scale), labels will correspond to much more specific areas of investigation in the geographic knowledge domain, e.g., "snowfall" or "redevelopment."

Cartography is essentially a science dealing with the transformation of spatial information (9). Following this paradigm, a number of geometric and topological transformations are applied to the raw geometric configuration produced by neural network training and, finally, symbolization occurs in off-the-shelf geographic information systems (GIS) software. This final step is driven by traditional cartographic considerations regarding visual hierarchies, here conveyed through color choices and manipulation of labels and line sizes. Label placement is performed automatically by GIS software.

Large-Format Knowledge Domain Visualization

With a display area of almost 12 square feet, the physical size of this visualization is more in tune with traditional cartographic output than snapshots presented in a journal paper (4) or

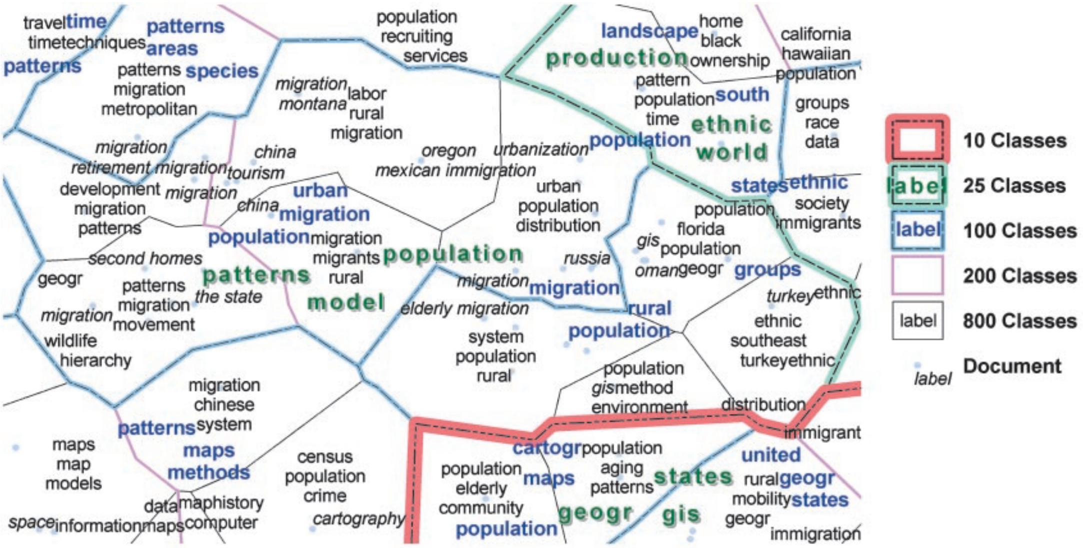

Fig. 2. Portion of a visualization of several thousand conference abstracts with simultaneous display of five cluster levels and individual documents.

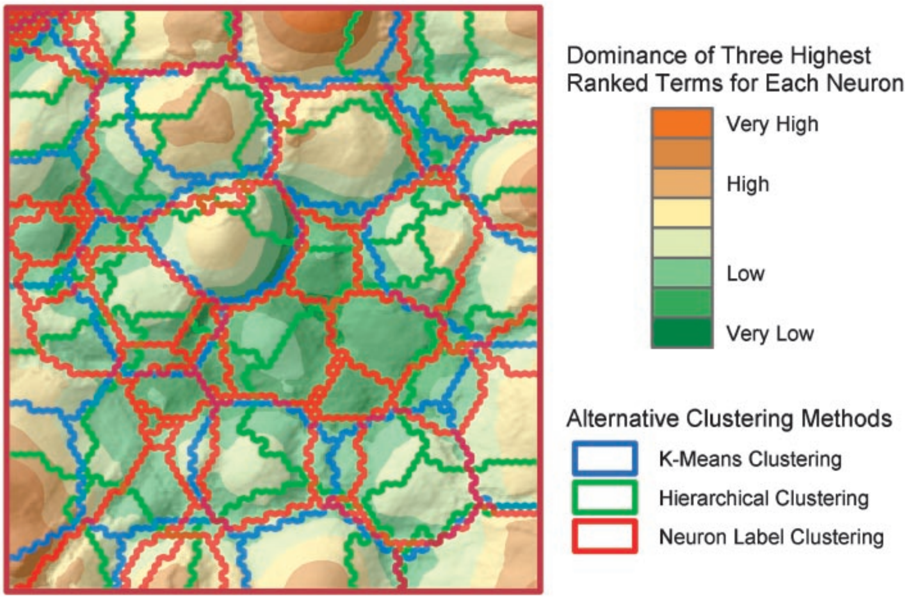

Fig. 3. Use of a term dominance surface to visually evaluate different clustering solutions.

interfaces heavily influenced by a limited screen size. It becomes possible to present multiple cluster levels simultaneously, making the use of hierarchical clustering particularly advantageous from a graphical point of view, because high-level cluster boundaries always also form lower-level boundaries. Rich labeling complements the extensive geometric structures created through this spatialization of conference abstracts, endowing the result with a remarkably map-like look (Fig. 2). A complete poster-size presentation of the result is available as Fig. 6, which is published as supporting information on the PNAS web site.

Creation of such large-format visualizations of knowledge domains is useful in various circumstances, especially in light of recent trends toward collaborative visualization (10). These efforts are complemented by a growing number of technologies that support the display of large-format visualizations, e.g., ImmersaDesk and DisplayWall. Interestingly, the major metaphors underlying the use of those technologies for visualization purposes, like drafting table or wallboard, correspond to traditional environments for cartographic map use.

Large displays on a static medium should not be easily dismissed either, especially when it comes to introducing novices to a knowledge domain and for establishing common ground among collaborating researchers. In those settings, these visualizations should be called “stable” rather than “static.” This has been one of the enduring qualities of large-format geographic maps. For example, when introducing proposed changes to a land-use ordinance in a town hall meeting, large-size maps are not merely used for illustration. Their purpose is also not to simply transmit an encoded geographic “message” and certainly not to gain insight into a phenomenon, as is the case for most scientific visualizations. Instead, these maps help to establish a shared frame of reference, without which human-to-human communication would be much more difficult.

Much work remains to be done to uncover the relative cognitive value of large-format visualizations in general, including those depicting knowledge domains. Similarly, it remains to be tested whether and under which circumstances static depictions are indeed inferior to highly interactive systems, as seems to be presumed by most knowledge domain visualizations.

Clustering Methods

In considering the use of clustering methods, it should first and foremost be pointed out that the purpose of clustering in this line of research is not to discover optimal feature space partitions. Instead, clustering serves as a stepping-stone in the support of visual exploration toward domain comprehension. Note that visual exploration does not necessarily imply interactivity in a human–computer interaction sense. Arguably, viewers of richly symbolized but static knowledge domain visualizations are engaged in a process of visual exploration as well.

The choice of hierarchical clustering to create the large-format visualization discussed earlier is driven by the advantages it offers graphically, conceptually, and computationally. Its nested structure makes simultaneous display of multiple cluster levels feasible (Fig. 2). At lower cluster levels, only truly new geometric elements have to be added, as long as the cluster hierarchy is properly conveyed through a visual hierarchy (e.g., use of line thickness to convey cluster level). However, certain problems associated with hierarchical clustering are also apparent. The nested structure comes at the cost of a suboptimal tessellation of the n-dimensional input space. For example, notice the appearance of similar labels near the peripheries of neighboring clusters (Fig. 2), indicative of the tension between a strict partitioning mechanism and the continuous nature of the self-organizing map.

The SOM method, with its field-like continuous conceptualization of a high-dimensional information space, makes exact partitioning indeed difficult, especially in transitional zones. It would be useful to know how well different clustering methods perform under these conditions. Apart from standard statistical approaches, e.g., an investigation of within- and between-cluster variances, it is possible to use spatialization to that end as well.

Visual and computational overlays of various thematic layers on the basis of a common coordinate system have been a mainstay of geographic information systems philosophy for over three decades, since such operations were first proposed in a precomputer setting (11). Similarly, one could overlay different clustering solutions onto the same neuron geometry, which is illustrated by Fig. 3 for three cluster solutions:

(*i*) a k-means solution ($k = 25$);

(*ii*) one level derived from a hierarchical clustering tree (at the 25-cluster level);

(*iii*) a solution based on a method we call neuron label clustering, in which neighboring neurons are merged into clus-

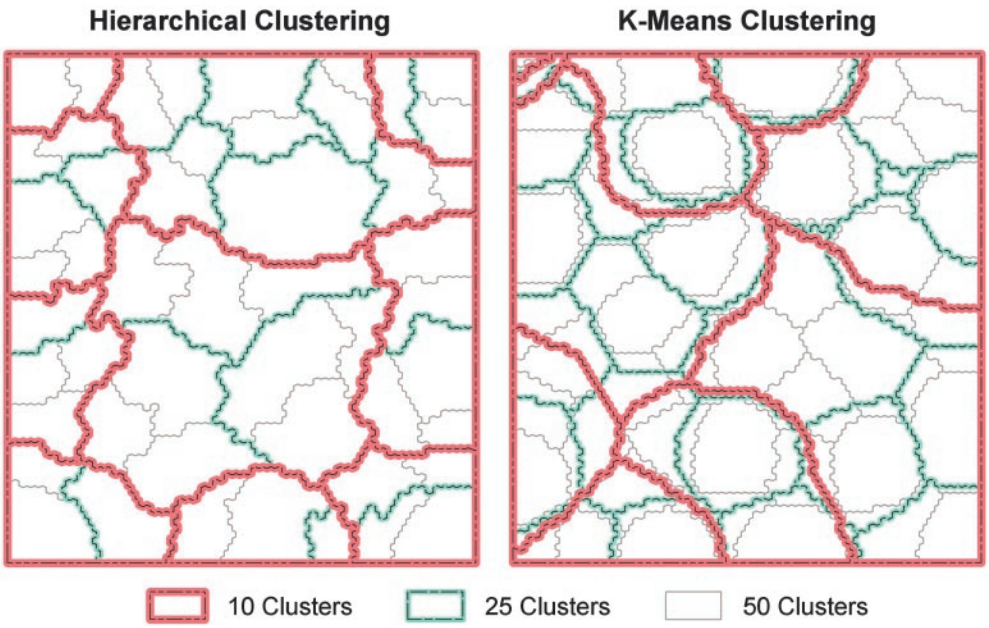

Fig. 4. Comparison of simultaneous display of multiple cluster levels based on two different clustering methods.

ters if their highest-weighted label terms are identical. This is similar to the clustering method proposed by Chen *et al.* (12).

Underneath, structures in the continuous information space are shown by means of a term dominance landscape, which expresses how dominant each neuron's top three label terms are with respect to all of the terms associated with a neuron. Because the training of SOM neurons is based on a dissimilarity/distance coefficient (in this case, the Euclidean measure), neighboring neurons will tend to have similar *n*-dimensional vectors associated with them, leading to a formation of extended mountain ranges. Higher "elevations," shown in brown tones, indicate a more coherently organized theme. Local minima may indicate a lack of distinct topical focus. "Clusters" incorporating those minima should thus be treated with some caution. Although superficially similar to other landscape-type knowledge domain visualizations, there are significant differences. Mountain ranges are formed by dominant combinations of keywords, i.e., major topics, across a large number of documents, which contrasts with a representation sometimes encountered of a majority of documents as local maxima (i.e., peaks) that seems to conflict with the continuous nature of the landscape metaphor. Formation of mountains is also not based on the density or number of documents that fall within its reach (13).

Valleys in the term dominance landscape correspond to transitional or overlapping topics between the dominant themes. This is again different from other landscape-type knowledge domain visualizations, where valleys mostly remain unpopulated by documents and must therefore be presumed to be void of meaning (13, 14).

Each clustering approach has distinct characteristics. Although the nested structure of hierarchical clustering has obvious advantages for graphic design and interaction, it has a tendency to cut through landscape features without obvious justification. The *k*-means method merges neighboring neurons into relatively evenly shaped and sized chunks, related to its use of the same objective function as the standard SOM training algorithm used here.

Of the three methods, the neuron label clustering approach matches the dominance landscape best, which makes sense because weighted term labels form the basis for computation of those two layers. Note how closely cluster boundaries follow "valley" features in the landscape, whereas "mountains" are enclosed. However, it offers the least control over granularity, which makes it difficult to create multiscale interfaces for exploration of knowledge domains.

Contrary to this, cluster levels in hierarchical and *k*-means clustering can be precisely chosen, as shown in Fig. 4. As mentioned earlier, the nested structure of hierarchical clustering reduces graphic and conceptual complexity (although we are not aware of human subject studies specifically investigating this issue). The *k*-means solutions appear graphically more complex, with plenty of overlapping clusters at different levels of *k*. On closer examination, some interesting observations emerge. Notice how some of the clusters at the 50-cluster level remain encircled and undivided by boundaries at the 25- and 10-cluster level, indicating agreement among different *k*-means solutions regarding these core areas. Interestingly, those cluster cores correspond to the major mountain ranges in the term dominance landscape (compare Fig. 3). On the other hand, peripheral clusters are formed at the 50-cluster level that are bisected at higher cluster levels. Those peripheral clusters correspond to either subtopics (i.e., divisions of larger topics), indicated by minor peaks within larger mountain ranges, or transitional/overlapping themes, shown as valleys in the term dominance landscape.

Compared to hierarchical clustering, the *k*-means method offers more optimal partitioning. On the other hand, it provides much better granularity control than neuron label clustering. Fig. 5 offers one suggestion for leveraging those characteristics while eliminating the complexity caused by cluster boundaries that do not coincide across multiple scale levels. The term dominance landscape is here combined with a labeled *k*-means solution, in which the cluster boundaries themselves are not shown explicitly but are at work in the background to automate placement of the computed cluster labels. Font size expresses the rank of a label term for that cluster. A semantic zoom operation is illustrated, during which a switch from a 25-cluster to a 100-cluster solution occurs. The mountain range labeled "population" is now shown in greater detail, breaking up related research topics into smaller categories, labeled "ethnic" to the right of the main peak and "migration" to its left. The location of these subcategories is significant, because the extensive use of

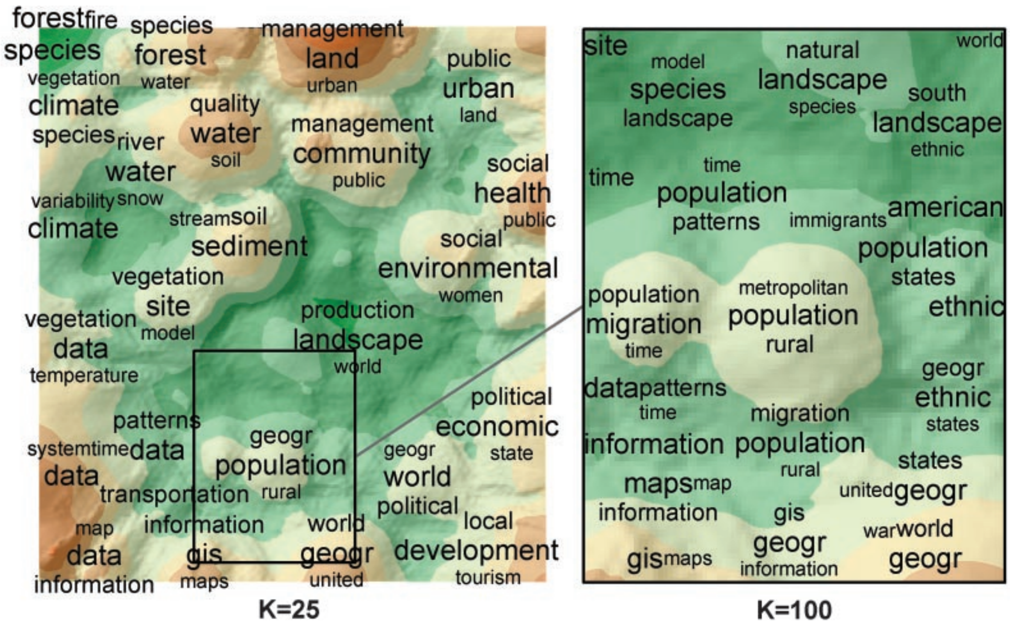

Fig. 5. Use of *k*-means clustering in combination with a term dominance landscape to support semantic zooming.

computational tools in migration studies warrants a position between the core population peak and the regions in the lower left of the global map focused on computational methodologies. This is quite different from studies of ethnic issues, which are typically grounded in a qualitative descriptive research paradigm, like many of the topics associated with the right half of this spatialization.

In summary, the purpose of clustering in knowledge domain visualization is not a provision of the "single best" and "true" partition of a domain, but rather one that may be useful under given circumstances. The examples discussed in this section demonstrate that the purpose of spatialization in the mapping of knowledge domains could extend beyond the creation of end-user tools. The computational procedures underlying multiscale visualizations may themselves be subject to visual inspection, and the resulting insights can inform the development of new or improved domain visualization methods.

Conclusion

This paper is largely driven by a desire to instigate reflection on the promise of the geographic metaphors and cartographic techniques that seem at the heart of so many knowledge domain visualizations. It raises important questions about the design of knowledge domain visualizations, such as: How far can we go in pursuit of cartographic metaphors? Is interactivity always necessary? Is there a role for static visualization in supporting discourse on the state and evolution of knowledge domains? Does the cognitive plausibility of certain visual approaches (e.g., a nested hierarchical structure) override a potential lack of computational plausibility? What would be the value of a convergence between knowledge domain visualizations and recent collaborative visualization developments?

This paper has demonstrated the possibility of creating large-format knowledge domain visualizations that emulate many aspects of traditional geographic depictions. Abstraction and scaling remain some of the most promising areas of cartographic influence on knowledge domain mapping efforts. In this context, this paper has presented an approach, informed by geographic information science, for the use of visual overlays to compare and validate different cluster techniques. The discussed techniques could of course be applied to similar data from other knowledge domains, as has been demonstrated elsewhere (15). We are currently developing a system aimed at providing a streamlined work flow for the creation of map-like knowledge domain visualizations. Future experiments involving both computational and human subject methodologies will help shed further light on the specific means for implementing useful map-like knowledge domain visualizations.

The research presented here is partially supported by the Louisiana Board of Regents Support Fund, Grant LEQSF(2002-05)-RD-A-34.

1. Card, S. K., Mackinlay, J. D. & Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think* (Morgan Kaufmann, San Francisco).
2. Skupin, A. (2000) in *InfoVis 2000* (Institute of Electrical and Electronic Engineers Computer Society, Salt Lake City), pp. 91–97.
3. Skupin, A. (2002) in *Visual Interfaces to Digital Libraries,* Lecture Notes in Computer Science, eds. Börner, K. & Chen, C. (Springer, Berlin), Vol. 2539, pp. 161–170.
4. Skupin, A. (2002) *IEEE Computer Graphics and Applications* **22,** 50–58.
5. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, T., Paatero, V. & Saarela, A. (1999) in *Kohonen Maps*, eds. Oja, E. & Kaski, S. (Elsevier, Amsterdam), pp. 171–182.
6. Lin, X. (1992) in *IEEE Visualization '92* (Institute of Electrical and Electronic Engineers Computer Society Press, Los Alamitos, CA), pp. 274–281.
7. Rushall, D. & Illgen, M. (1996) in *InfoVis 1996* (Institute of Electrical and Electronic Engineers Computer Society Press, Los Alamitos, CA), pp. 100–107.
8. Salton, G. (1989) *Automated Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison–Wesley, Reading, MA).
9. Tobler, W. (1979) *Am. Cartogr.* **6,** 101–106.
10. Brewer, I., MacEachren, A. M., Abdo, H., Gundrum, J. & Otto, G. (2000) in *InfoVis 2000* (Institute of Electrical and Electronic Engineers Computer Society, Salt Lake City), pp. 137–141.
11. McHarg, I. (1969) *Design with Nature* (Natural History Press, Garden City, NY).
12. Chen, H., Schuffels, C. & Orwig, R. (1996) *J. Visual Commun. Image Rep.* **7,** 88–102.
13. Boyack, K. W., Wylie, B. N. & Davidson, G. S. (2002) in *Visual Interfaces to Digital Libraries*, Lecture Notes in Computer Science, eds. Börner, K. & Chen, C. (Springer, Berlin), Vol. 2539, pp. 145–158.
14. Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. & Crow, V. (1995) in *InfoVis 1995* (Institute of Electrical and Electronic Engineers Computer Society, Atlanta), pp. 51–58.
15. Börner, K., Chen, C. & Boyack, K. W. (2003) in *Annual Review of Information Science and Technology*, ed. Cronin, B. (Information Today, Inc., Medford, NJ), Vol. 37, pp. 179–255.