

Learning Geoscience Categories *In Situ*: Implications for Geographic Knowledge Representation

Boyan Brodaric
bmb184@psu.edu

GeoVISTA Center, Penn State
Geography
302 Walker Building
University Park, PA 16802 USA
814-865-3433

Mark Gahegan
mark@geog.psu.edu

ABSTRACT

This paper explores the development of categories shared in the field logging of a region by a team of geologists. Visualization, neural networks and spatial statistical tools are employed to gain insight into the complex space of attributes observed, and into the categories developed. Background material and a discussion of results examines the findings in the light of research into category development, and specifically how categories are thought to be formed and modified as part of the (geo)scientific process and the situations encountered. Results show that (1) category discrepancy exists between individuals; (2) category development or revision is evident among individuals; and (3) that some categories do not seem to be totally defined by observed data alone. The results imply that contextual factors should also be considered when adopting ontological approaches to information representation.

Categories and Subject Descriptors

H1.1 [Systems and Information Theory]: Value of information.

H1.2 [User/Machine Systems]: Human information processing.

H2.8 [Database Applications]: Spatial Databases and GIS.

H3.3 [Information Search and Retrieval]: Clustering.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright 2001 ACM XXXXXXXXX/XX/XX ...\$5.00. This copy is posted by permission of ACM. The original is at <http://www.acm.org/dl/>.

Brodaric, B. & Gahegan, M. (2001). Learning Geoscience Categories *In Situ*: Implications for Geographic Knowledge Representation. *In Proceedings, ACM-GIS 200: The Ninth ACM International Symposium on Advances in Geographic Information Systems, Atlanta, GA, Nov. 9-10, 2001.*

General Terms

Measurement, Reliability, Experimentation, Human Factors.

Keywords

Category development, situated learning, information ontology, self-organizing maps, classification, geological fieldwork.

1. INTRODUCTION

Field scientists are often the originators or arbitrators of categories used in the production of maps. In the geosciences, categories may be developed *in-situ* by teams of field scientists working together in a coordinated fashion. Such categories are not necessarily fixed at the outset, nor are they perfectly shared between all the individuals concerned. Rather they develop as a result of experience, exposure, training and other factors that are difficult, if not impossible to control for. The categories in geological mapping differ somewhat from those in other forms of mapping in that geologists form categories in the process of constructing a consistent geoscientific model to explain the evolution of a region. Because geoscientific environments are typically *open* systems characterized by compositional heterogeneity and stochasticity, and because gaps in evidence and theory are abundant [16], and observations and interpretations somewhat subjective, it is possible for multiple models (i.e. maps) to validly explain an area [22] even when categories are shared.

This paper sets out to study the development of geoscientific categories in such environments. The logging records of individual field scientists are analyzed for consistencies and discrepancies. Because of the large number of variables and overall complexity of the data, geocomputational methods are employed for this analysis to gain insight into the nature the categories and their development through time. Previous work has shown that some categories do not appear to be completely definable in terms of their attributes alone, but in addition require significant head knowledge from the geologist that is *not* captured in the data [4]. Needless to say, this apparent gap is detrimental to the effective and reliable use of the collected data and resultant interpretations. We investigate this gap to gain insight into the geologic categorization process and associated computing implications.

This work also forms part of a longer-term project examining space-time information semantics. Specifically, we consider whether ontologic approaches to geospatial knowledge representation should be augmented by situational knowledge.

Ontologic approaches to geospatial information semantics [e.g. 3, 7] provide meaning for a geospatial object via the taxonomic (conceptual) and logical (theoretic) structure being applied—i.e. its *perspective*. For instance, ecologic and geologic perspectives of a body of rock would differ radically. However, it is also probable that the same ontologic perspective could satisfy multiple models; i.e. geologists committed to a specific categorical scheme, might still generate different maps for the same area.

To help explain such variation, research in category theory suggests that categories, and thus ontologic structure and perspective, are affected by the specific *situations* in which they are created and used [23]. For information semantics this implies that the meaning of a geospatial object might be contextual, in part due to ontologic commitments, and in part due to epistemic considerations [21] (i.e. how knowledge is acquired [18]) involving the situations leading to the development of the categories, and the situations (or models) arising from the assignment of the categories to specific geospatial entities. In the realm of geologic mapping such situational effects might translate into the categories being viewed differently by individuals depending on the specific terrain encountered, the sequence of observation, their prior knowledge and expertise, as well as their methodological orientations, etc.

This investigation is primarily conducted to test the hypothesis that field-based geoscientific cognition is situated; if this is so, then the need to weave situations into knowledge representation frameworks is substantiated. The field data of three geologists are compared and contrasted using geocomputational methods. These geologists worked as a team to develop common ontology (categories) and a common model (map), using identical computer-based data recording technology, data structures, vocabulary, and field methods [4]; in short they had common ontological commitments and methodologies, and we investigate the similarity of their epistemic orientations by comparing and contrasting the development of individuals' categories, through time. Differences between and within individuals would signify varied understanding of the categories, and require greater capture of the context in which the categories were created and applied.

This paper is organized as follows: Section 2 describes the study area and cognitive background; Section 3 details the methods used to associate categories and data; Section 4 reports on the results; Section 5 discusses some implications of the results, and Section 6 closes with a brief conclusion.

2. Background

2.1 Study Area and Data

The study data comes from a geological mapping project conducted by the Geological Survey of Canada, and various other partners, in northeastern Canada (Figure 1) [12]. The data consists of six 1:50,000 scale geological maps (60x80 km) [27], several categories (interpreted rock formations) used to label map areas, and a large volume of supporting evidence gathered in the field. The data were developed by geologists during the 1998-99

field campaigns, when they were daily transported to areas of interest for note taking. In the evening the field notes were assimilated into a common geospatial database at the shared base camp, where analysis was conducted to aid map construction, guide future work, and synchronize ongoing interpretations, including the evolving categories, models and map. The individuals thus made every effort to maintain semantic consistency in the field. After each field season the evolving interpretations (but not evidence) were published in various scientific outlets [e.g. 12, 27].

We obtained the complete data after the 1999 results were published, and analyzed a subset of this data (i.e. data from the three main geologists, including the 6 maps, 1535 sites and 16 categories). This subset was chosen for its large sample size and even thematic and geographic distribution; e.g. at least 2/3 of the polygons were visited by two or more individuals for significant cross-referencing of evidence. The subset mainly contained multiple descriptions at a site for rock composition (rock type, minerals, textures, etc.) and spatial disposition (orientations of planar and linear fabrics, etc.); each description also possessed a date-stamp and an enumeration of its categorical significance, as determined by the geologist while in the field.

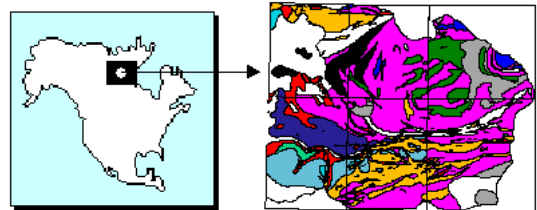


Figure 1. The study area and resultant geological maps.

2.2 Ontology, Concepts, Theories, Situations

The geologic categories developed by the individuals represent an ontology, inasmuch as the taxonomic and logical structure of the categories provides meaning to the observed phenomena [19]. They are however rather more unique than some philosophic categories [25] or geospatial categories [8], in that they may derive from a specific knowledge acquisition exercise [5]. Guarino [11] captures this decrease in generality in his three-tiered approach to information ontology, where broadly accepted ontologies occupy a top-level, less generic domain-specific ontologies occupy a domain and task level, and specific implementation ontologies occur at the application level. It is likely that highly differentiated ontologies, such as those at the application level (e.g. ontologies for geologic surveying), are subject to epistemic influences, and this approaches a situated cognition position in which application ontologies develop ‘on the fly’ as part of a cognitive process [6].

Traditional approaches to categorization do not incorporate such dynamism. In the *classical* view, concepts are uniquely determined by defining attributes, and associated necessary and sufficient rules, but classical categories are not well supported empirically [17] (for an alternate view see [26]). Consequently, in

the *probabilistic* view concepts are organized by a graded concentration of characteristic attributes, causing some examples to be more typical of a concept than others; but this only accounts for some but not all categories [17]. In the *theory* view, categorization is largely due to theories that are held by people to explain the world [17]; but people reason and act differently in different situations and cultures [23]. In summary, classical, probabilistic and theory-driven views on their own fail to account for how concepts are developed and used in response to different situations, and this challenges the notion that meaning is solely defined in terms of stable concept definitions and taxonomic structures, such as those on which most ontologies are based.

Dynamic approaches to cognition indicate that conceptual structures may be constructed or revised according to the situation experienced [2], or function performed [24], but it is uncertain how much the mind [1], body [15], or the environment [6] are mainly responsible. All sides of this debate suggest that human behavior and conceptual structure evolve, at least partly based on what prior physical, social, and conceptual interactions occur, and what motivational and environmental influences are brought to bear [23]: e.g. category development may trend from probabilistic to theoretic as experience and expertise increases [24, 28]. Understanding a particular geoscientific category, object, or action, may therefore require understanding of the accrued set of inferences and situations that led to it, in addition to the situation at hand. From the viewpoint of geoscientific data this might include the methods, actions, intentions, social interactions, and various data sources used to generate the data, as well their historical development. In short, geoscientific meaning may be historical and contextual, as well as data-oriented, taxonomic and theoretic, but these claims require more empirical investigation.

3. Analysis and Visualization Methods

This study investigates the nature of field-based geoscientific category development. It examines the association between field observations and derived categories, generated in the field by geoscientists as a standard part of their geologic mapping exercise. These data were obtained and analyzed by us, months after the actual mapping was completed. Geostatistical methods were applied to the acquired data to track category development over time in a multi-dimensional feature space, and the Self-Organizing Map neural network (SOM) was used to simulate categorization, and investigate the separability and similarity of the categories. Significant effort went into the conversion of the field data from textual and nominal form to numeric linear vectors suitable for usage in a multi-dimensional ‘feature’ space. The details of this conversion are described in [4]; of significance is the fact that optimal analysis was achieved when the input vector dimensions were weighted according to geologic theory and adjusted to emphasize critical dimensions determined by the geologists while in the field. The high dimensionality of the input vector space (> 100 attributes) and the complexity of the data warranted the choice of neural analysis techniques [9].

3.1 Self-Organizing Map

The SOM [13] is an example of an unsupervised inductive classification or clustering method. The SOM consists of a matrix of prototype nodes that are adjusted by repeated application of sample data: i.e. differences between the sample data and the nearest prototype (calculated via a distance measure—in this case Euclidean) are propagated to a locus around that prototype. SOMs can be labeled with categories after training and are thus a good measure of confusion, as nodes may possess multiple category labels [14]. Also, since the SOM reflects the topology of the input data, its visualization provides an indication of the spatial configuration of feature space, though actual distances are not preserved. The SOM is described in greater detail in [13].

Supervised SOM categorization (i.e. learning vector quantization [14]) is very similar to the unsupervised case: a set of initial prototype nodes possessing category labels are scattered in feature space. Input samples find the nearest node, which is moved toward the sample if categories match, and away from the sample if categories conflict; in general there is no propagation of adjustment to other nodes. Accuracy values for a set of samples can be determined by applying the data to a trained SOM and recording the proportion of correctly identified samples.

3.2 Sammon Mapping

The Sammon mapping [20] is a form of multi-dimensional scaling which transforms an n -dimensional feature space into two dimensions suitable for visualization. The Sammon mapping preserves relative distances between vectors, but topology is not preserved. It is used in this study to visualize the field data and SOM, for visual comparisons of category similarity.

3.3 Median Minimum Distance

The Median Minimum Distance (MMD) is a measure of central spatial tendency [14]. When used to compare two clusters of vectors, it effectively measures the distance between cluster centroids, indicating their degree of similarity. When used to compare one cluster to itself, the resulting distance can be viewed as a measure of the diameter of the central tendency. It is calculated by taking the mean of the medians of the shortest distances between vectors. Due to asymmetries, the MMD of two vector sets may be averaged (as it was in this study).

4. Results and Discussion

4.1 Expertise

Training the SOM to partition the field data into discrete category clusters resulted in relatively low overall accuracy when considering either individual geologists or all geologists together (Table 1). This implies that the categories could not be inductively determined from the data alone, as their attributes do not form a unique configuration within feature space (Figure 3). It also concurs with past results that suggest the determination of categories from geoscientific field data is in part *probabilistic*, and in part reliant on other expertise (such as *theories*) [4].

Table 1. Results of supervised SOM classification on all (16) and some (3) categories.

Category	sites	nodes	cycles	Accuracy%
ALL	1535	300	15000	37.85
ALL	1535	500	25000	55.31
ALL	1535	700	35000	59.35
ALL	1535	900	45000	59.87
ALL-geol 1	563	500	25000	66.61
ALL-geol 2	554	500	25000	55.78
ALL-geol 3	418	500	25000	61.48
A, B, C	959	200	10000	80.60
A, B, C	959	350	17500	79.25
A, B, C	959	500	25000	81.96

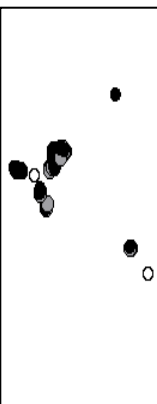


Figure 3. Sammon view of categories A, B and C (black, gray, white). Note that similar data are grouped together leading to multi-modality and overlap of the categories.

4.2 Situation and Individual Bias

Comparison of individuals' data within several categories demonstrated quite a range of similarity and difference. Figures 4a and 4b depict this range in that individuals appraised category A (Figure 4a) relatively differently, and category B (Figure 4b) relatively similarly, according to their distributions of field data in feature space. However, in neither case could individuals' data be considered fully congruent. This range of categorical (dis)similarity was repeated in other categories, and concurs with our previous study of a single category [4]. Also, similarity was sometimes, but not always, correlated to geographical proximity of observations between individuals, suggesting exposure to similar physical situations may have increased categorical congruity in some cases.

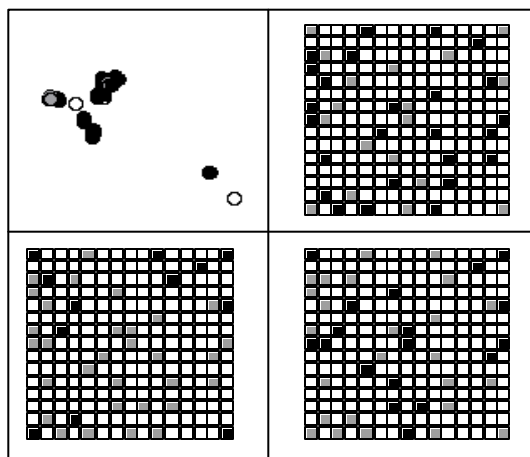


Figure 4a. Feature space view of category A. The upper left Sammon view depicts the placement in feature space of three individual's data as white, gray and black symbols; the remaining panels show one individual's data (black) contrasted with the remaining two individuals' data (gray) as manifest within the unsupervised SOM.

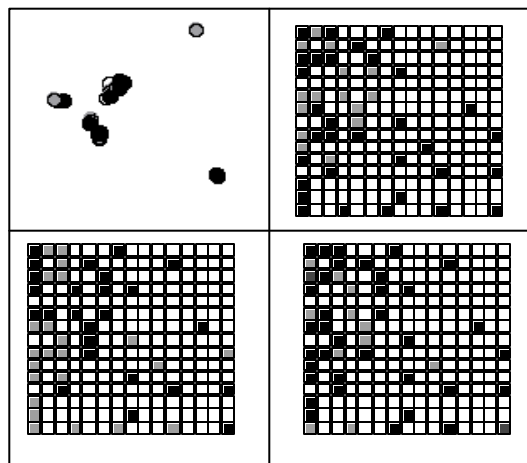


Figure 4b. Feature space view of category B. See Figure 4a for a description of the panels.

4.3 Learning

Individual Learning: Category development in individuals was investigated by tracking the size of a category's central tendency in feature space over time (Figure 5, left hand side). Reduction in size represents clustering or category refinement and is consistent with *probabilistic* categorization inasmuch as clustering denotes a trend to core attributes (Figure 5, left, category B); increase in size represents dispersion or category revision and may indicate categorical uncertainty (Figure 5, left, category C, dashed line); a static pattern indicates category stability and possibly acquired expertise (Figure 5, left, category C, dotted line); and a fluctuating pattern may indicate categorical uncertainty, limited expertise, or situational response (Figure 5, left, category C, solid line).

In addition, two phases of category development were observed; a short initial phase marked by rapid changes of the position of the category within feature space (ascent or descent), perhaps related to initial *probabilistic* exploration of the category, and a longer, more regular, second phase in which the category is refined, perhaps guided by *theory* developed in the first stage.

In summary, probabilistic categorization is suggested in some categories, but not in others, and some are well formed at the outset. Support for uncertainty, or varying expertise, in specific categories is also apparent, and this uncertainty appears to vary between and within some individuals over time.

Group Learning: Pairwise similarity of individuals within categories was also compared over time (Figure 5, right hand side). Both convergence and divergence of individual views of a category was observed, though some level of difference between individuals was consistently apparent. Moreover, no clear pattern of similarity between individuals emerged across categories, as similarities between pairs of individuals in one category were not necessarily replicated in other categories. Perhaps the most telling aspect of these particular results is the implication that individuals maintained strong personal categorical trends in the face of varying situation and pooled group experience.

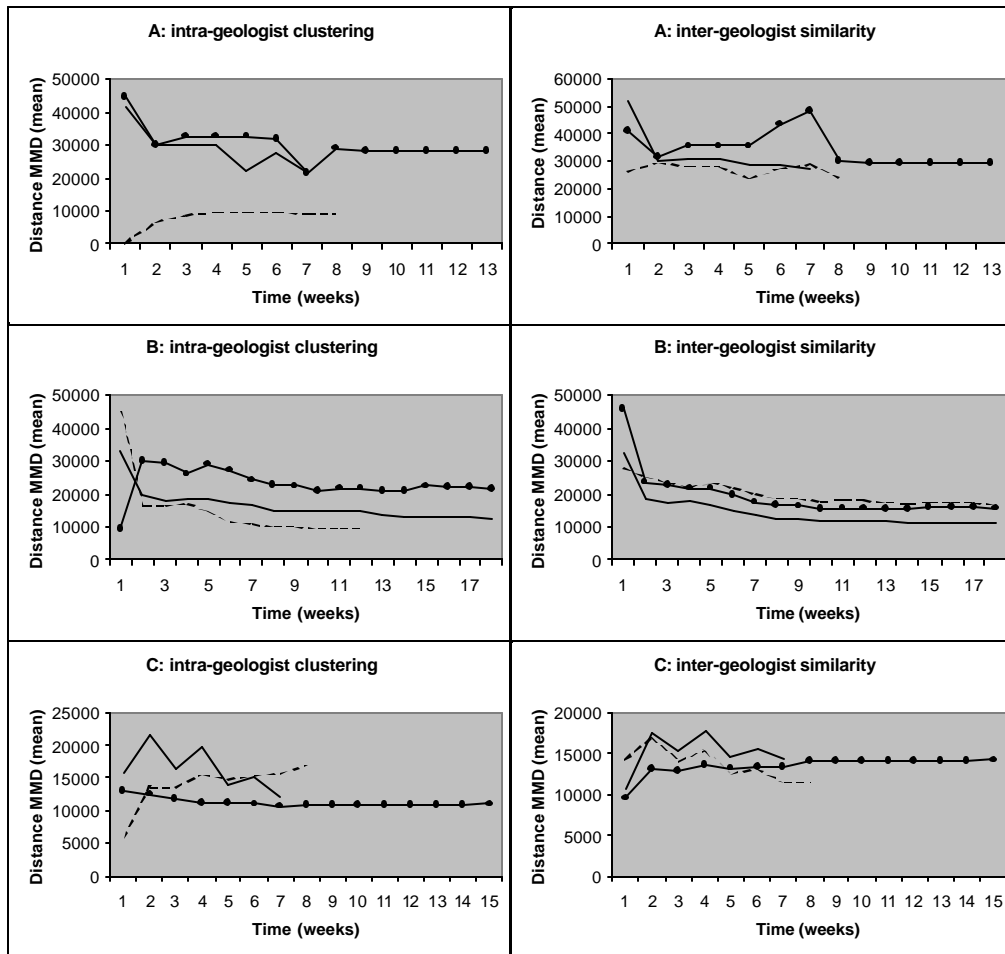


Figure 5. Individual learning (left) and group learning (right) in categories A, B, C. The charts on the left represent the median diameter of an individual’s category in feature space calculated weekly for distinct categories. Reduction of this distance indicates clustering to a focal area. The charts on the right represent the distance between individuals’ categories in feature space, calculated weekly for distinct categories--each line represents a pairwise comparison of individuals. Note that high convergence is not attained, implying the presence of intra-categorical differences between individuals.

5. Discussion

Of particular significance are the results demonstrating that categorization involved (1) factors extrinsic to the data and (2) personal trends within categories that varied in time. These results imply that ontologic consistency is difficult to maintain because individual views of common categories may differ amongst and within individuals, even when controlled for. This further suggests that understanding, using and transforming geospatial information may require more than ontologic knowledge, it may require sufficient context such as uncertainty, error, theory, history, usage, and other epistemic factors that describe how geoscientific categories are derived from situations. Such factors contribute to explaining how categories come to be, and how they are applied, and their inclusion in knowledge structures would seem to be necessary to understand and use categories that are dynamic and situated, such as those investigated in this study.

6. Conclusions and Future Work

As a general observation, the combination of statistical analytic, inductive machine learning and visualization techniques employed here provide very useful insight into category development, at least from the perspective of the data analyzed. Consequently a larger array of such methods will be applied in future work [10].

That the collected data does not completely differentiate the derived categories is not in itself surprising, but it is nonetheless a useful point from which to begin to identify other sources of information and knowledge, how they are used, and what effects they have on the mapping process and the reliability of the result.

The portrait of field-based geoscientific activity that emerges from these results is one of situated scientific cognition. Geoscientists bring their experience and knowledge to the field where the environment's capacity for infinite permutation requires of them constant adaptations of conceptual and pragmatic knowledge. It would seem that geoscientific ontology and phenomena maintain a

significant interrelation suffused with history and context. This argues for the inclusion of situational influences into ontologic knowledge structures to provide context to information. We are currently developing such structures and will be soon testing them with geoscientific data.

7. Acknowledgements.

Our thanks go to Dr. M. Takatsuka, Scott Miller, and Rob Harrap for their help in this work. This research is sponsored under NSF grant EIA-9983445 and supported by the GSC and USGS.

8. References

- [1] Arbib, M.A., and Hesse M.B. (1986). *The construction of reality*. Cambridge University Press, New York.
- [2] Barsalou, L.W. (1983). Ad-hoc categories. *Memory and Cognition*, 11, 211-227.
- [3] Benslimane, D., Leclercq, E., Savonnet, M., Terrasse, M.-N., and Yetongnon, K. (2000). On the definition of generic multi-layered ontologies for urban applications. *Computers, Environment, and Urban Systems*, 24, 191-214.
- [4] Brodaric, B., Gahegan, M., Takatsuka, M., and Harrap, R. (2000) Geocomputing with geological field data: is there a ghost in the machine? *Proceedings, GeoComputation 2000*, Chatham, August 23-25.
- [5] Clancey, W.J. (1993). The knowledge level reinterpreted: modeling socio-technical systems. *International Journal of Intelligent Systems*, 8, 33-49.
- [6] Clancey, W.J. (1997). *Situated Cognition: on human knowledge and computer representations*. Cambridge University Press, New York.
- [7] Fonseca, F. T., Egenhofer, M. J., Clodoveu, A. D. Jr., and Borges, K. A. V. (2000). Ontologies and Knowledge Sharing in Urban GIS. *Computers, Environment, and Urban Systems*, 24(3), 251-272.
- [8] Frank, A., and Raubal, M. (1999). Formal specifications of image schemata - a step towards interoperability in geographic information systems. *Spatial Cognition and Computation*, 1 (1), 67-101.
- [9] Gahegan, M. (2000). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, 32(1).
- [10] Gahegan, M, Takatsuka, M., Wheeler, M. and Hardisty, F. (2000). GeoVISTA Studio: A Geocomputational Workbench. *Proceedings, GeoComputation 2000*, Chatham, Aug. 23-25.
- [11] Guarino, N. (1997). Understanding, building, and using ontologies. *International Journal of Human-Computer Studies*, 46, 293-310.
- [12] Hanmer, S. and Relf, C. (2000) Western Churchill NATMAP Project: new results and potential significance; *Proceedings, GeoCanada 2000*, Calgary, May 29-June 2.
- [13] Kohonen, T. (1997). *Self-organizing maps*. Berlin, NY.
- [14] Kohonen, T., Hynninen, J., Kangas, J., and Laaksonen, J. (1995). *SOM_PAK, The Self-Organizing Map Program Package, version 3.1* (April 7, 1995).
- [15] Lakoff, G. (1987). *Women, Fire and Dangerous Things*. University of Chicago Press, Chicago.
- [16] Mann, C.J., 1993. Uncertainty in Geology. In Davis, J.C. and Herzfeld, U.C. (Eds.), *Computers in Geology—25 Years of Progress*. Oxford University Press, New York.
- [17] Medin, D.L. (1989). Concepts and Conceptual Structure. *American Psychologist*, 44(12), 1469-1481.
- [18] Peuquet, D., Smith, B., and Brogaard, B. (1999). *The Ontology of Fields. Report of a Specialist Meeting held under the auspices of the Varenus Project, 11-13 June 1998, Bar Harbour, ME*.
- [19] Raper, J. (1999). Spatial Representation: the scientist's perspective. In P.A. Longely, M.F. Goodchild, D.J. Maquire, and D.W. Rhind (Eds.), *Geographical Information Systems: Principles and Technical Issues*, Wiley, New York.
- [20] Sammon, J.W. Jr. (1969). A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18, 5, 401-408.
- [21] Schuurman, N. (1999). Critical GIS: Theorizing an Emerging Science. *Cartographica*, 36(4).
- [22] Schumm, S.A. (1991). *To Interpret the Earth: Ten ways to be wrong*. Cambridge, New York.
- [23] Smith, L. B. and Samuelson, L. K., (1997). Perceiving and remembering: Category stability, variability and development. In Lamberts K. and Shanks, D. (Eds.), 1997. *Knowledge, Concepts, and Categories*. MIT Press, Cambridge, MA, 161-196.
- [24] Solomon, K. O., Medin, D. L., and Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences*, 3(3), 99-104.
- [25] Sowa, J.F. (1995). Top-level ontological categories. *International Journal of Human-Computer Studies*, 43, 669-685.
- [26] Sutcliffe, J. P., (1993) Concept, class, and category in the tradition of Aristotle. In Mechelen, I. V., Hampton, J., Michalski, R. S., and Theuns, P. (Eds.) *Categories and Concepts: theoretical views and inductive data analysis*. Academic, New York, 35-66.
- [27] Tella, S., Hanmer, S., Sandeman, H. A., Ryan, J. J., Hadlari, T., Mills, A. and Kerswill, J. A. (1999) *Geology, Parts of MacQuoid-Gibson lakes area, Kivalliq Region, Nunavut; Geological Survey of Canada, Open File 3701*; Geological Survey of Canada, scale 1:50 000.
- [28] Wisniewski, E.J. & Medin, D.L. (1994). On the interaction of theory and data in concept learning. *Cog. Sci.*, 18, 221-281.