A Crowd-Sourced Taxonomy for the Common-Sense Geographic Domain

David Mark[1], Alexander Klippel[2], Jan Oliver Wallgrün[2]

[1]NCGIA & Department of Geography, University at Buffalo, NY 14261 USA
Email: dmark@buffalo.edu

[2]Department of Geography, The Pennsylvania State University, PA 16802 USA
Email: {klippel,wallgrun}@psu.edu

1. Introduction

How are geographic terms, concepts, and their referents related to each other? Is the so-called *geographic domain* a natural partition of reality, or, as some have suggested, is it just an *ad hoc* collection of things that geographers happen to be interested in? These questions are relevant to the various sciences that deal with geographic information.

Taxonomies and ontologies of the commonsense geographic domain were identified as key research goals for Naïve Geography (Egenhofer and Mark 1995) and are central to research on ontologies and semantics in general (eg. Janowicz et al. 2010). In this paper, we present a first step toward a commonsense taxonomy of the geographic domain, derived from a synthesis of behavioral studies of members of the American English-speaking general public.

2. Methods

2.1 Selection of Terms

Smith and Mark (2001) developed norms for geographic entities by conducting free-listing experiments, asking undergraduate participants to list examples of geographical things; 6 different phrasings of the question were used, and 373 participants provided examples. Participants listed terms for 15 seconds, and provided an average of 5.6 examples each. Together, participants provided 327 different terms. The most frequent example was "mountain" (f=224; 60% of participants).

In the current study, we gave participants 53 terms in a category construction task. These included the 36 most-frequent terms from the Smith and Mark (2001) elicitation norms, plus 17 additional terms with lower frequency, arbitrarily selected from that list. The exact instructions were: "*Please sort on how similar the things are that the words refer to.*"

2.2 Participants

The category construction task was administered using CatScan (Klippel et al. 2013) and participants were recruited via Amazon Mechanical Turk (AMT). One hundred participants took part in the experiment (54 female, average age 35.38). Participants received $1 + $.25 for their participation. All participants were native English speakers (we excluded 3 participants with a different language background, and one for indicating his age was 100). All participants live in the US. Mean grouping time was 8 min. Participants created on average 6.68 groups.

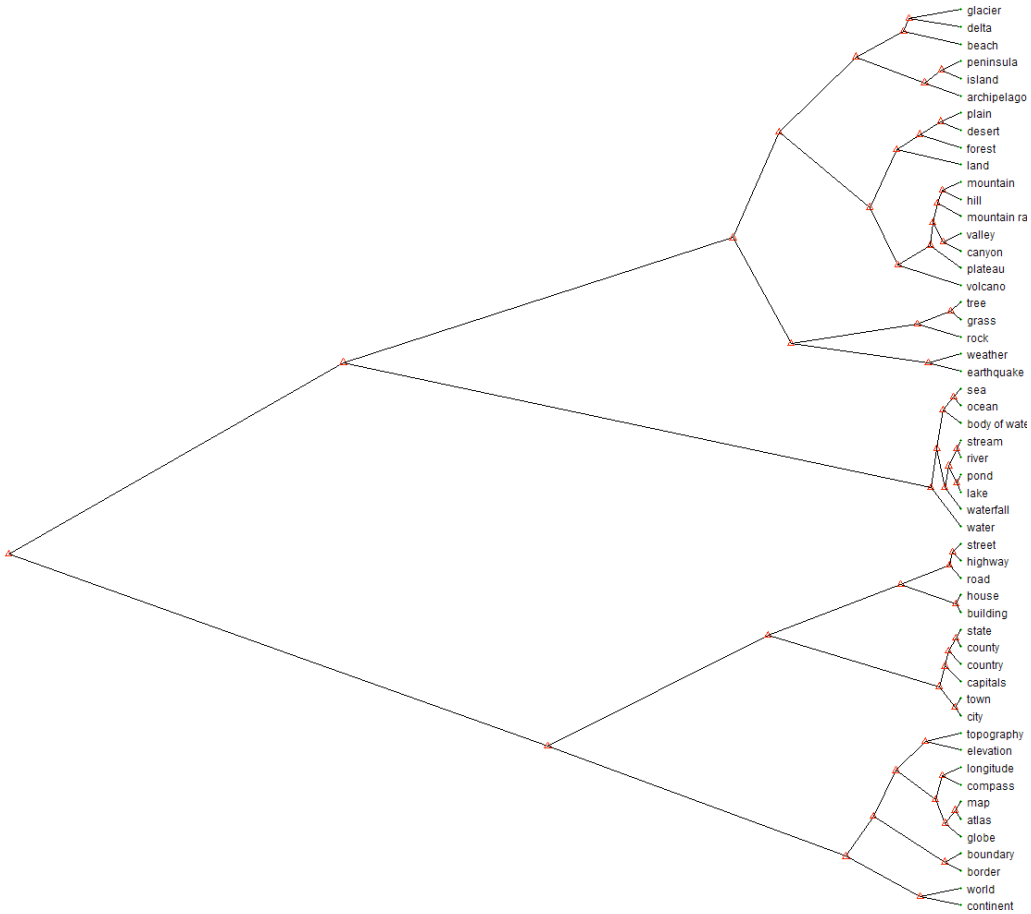

Figure 1. Taxonomy derived from Ward's cluster analysis.

3. Results

3.1 Interpretation of the Clusters

Figure 1 shows the taxonomy of geographic terms using (exemplarily) Ward's method. It may help our understanding of the dendrogram to use the terms and concepts of taxonomy in biological systematics. One key concept is the *clade*. In biological taxonomy, a *clade* is a group consisting of an ancestor and all its descendants. A *monophyletic* group includes one clade. A *polyphyletic* group includes more than one clade. Of course we realize that this is just an analogy, and that groups of geographic terms do not descend from common ancestral terms. Cluster validation (see section 3.2, below) identified 12 clusters that were robust across all clustering methods. The 8 clades below were considered to be monophyletic if they contain exactly one validated cluster, and polyphyletic if they contain 2 or more validated clusters.

Eight 'clades'. By 'cutting' the dendrogram in Figure 1 at an appropriate level, it can be divided into eight clades (sub-trees). For the purpose of structuring the discussion, we refer to the upper five clades as containing relatively natural entity types, and the lower three as containing less natural types (artificial/man-made, abstract concepts). Three of the more natural clades appear to be conceptually homogeneous (which we interpret as analogous to *monophyletic* clades in biology). There is a group that includes all of the water features and only water features. Another

clade is composed of small (sub-geographic) environmental entities (tree, grass, rock), and a third clade contains two types of dynamic entities. We believe that these three clades are uncontroversial. The other two 'natural' clades are heterogeneous (analogous to polyphyletic clades). One of the clades contains two clusters, each of which appears to be homogeneous: seven landform types, and four ecoregion types. The other polyphletic natural clade also contains two validated clusters: three shore-bounded land feature types, and entity types not very similar to any other terms in the study, although they do have a water component: glacier, delta, and beach.

Two of the three clades composed of less 'natural' entity types appear to have conceptual homogeneity. One clade includes components of the *built environment*, and has two coherent sub-clades: street-highway-road, and house-building. A second less-natural clade includes fiat administrative regions (state, county, country) and settlement types (town, city). The remaining clade on the less natural side is deeply polyphyletic, but contains three validated clusters. Four terms denote manipulable artifacts with a geographic purpose: compass, map, atlas, and globe. The placement of the term 'longitude' within this group is strange, but the validated cluster also includes topography and elevation. Another validated cluster within this clade consists of two terms: boundary and border. The last validated cluster within this clade includes world and continent. Some people, including an anonymous reviewer, feel that these terms should be on the 'natural' side of the dendrogram. However, at least in English, 'world' is not a synonym for 'earth', but is more conceptual. 'Continent' also may be thought of as more like a fiat object. But we note that the placement of the world/continent group within the dendrogram is surprising and deserves further scrutiny.

3.2 Cluster Validation

To corroborate this finding statistically, we developed a cluster validation technique referred to as cross-method similarity index (CMSI, Wallgrün et al. 2014). Figure 2 shows that the two-cluster solution mentioned above is the most stable conceptualization of the terms used in this experiment. In a nutshell: A CMSI index of 1 indicates perfect correspondence across 100 random samples and three different clustering methods.

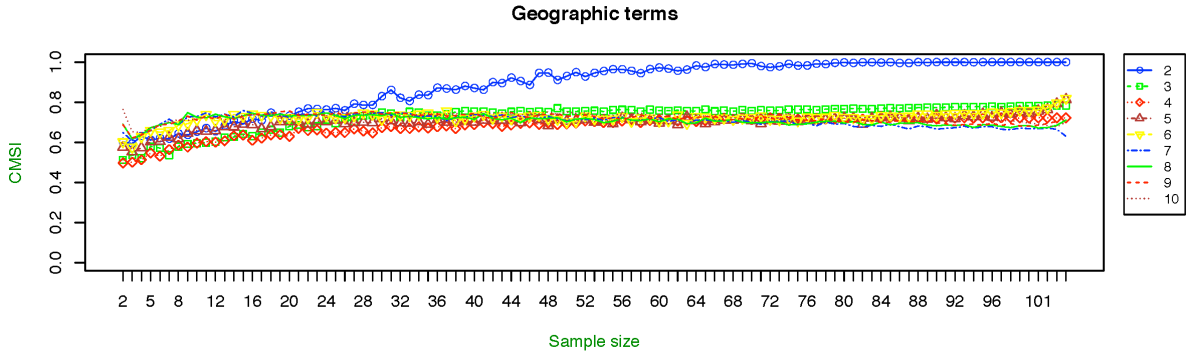

Figure 2. CMSI method for cluster validation. An index of 1 indicates a perfect correspondence across three cluster methods. Each sample-size-value is the average of 100 random samples.

To further understand the conceptual structure (taxonomy) of the terms employed in the experiment, we used an algorithm (Wallgrün et al. 2012) that analyzes tree structure of dendrograms (rather than cluster membership) and compares different clustering methods to validate the results. This approach identified 12 groups of terms that are robust across three

clustering analyses (Ward, average and complete linkage). As noted above, each of these 12 groups has a high degree of internal semantic-coherence, i.e., they appear to make sense.

4. Conclusions and Future Work

4.1 Artifacts

As noted above, one of the most obvious and clear result is that many participants separated natural geographic entity types from artificial ones. Artifacts apparently present a thorny problem for formal ontologies (cf. Borgo and Vieu 2009), and this presumably will present a challenge for integrating this crowd-sourced taxonomy into formal ontologies.

4.2 Future Work

Our next step in this investigation will be the strategic addition of more terms, to confirm or test the over-all conclusions and to fill in semantic gaps. Examples of the kinds of terms we wish to add include: some small non-geographic-related artifacts such as chair, book; some small animals; some very large mobile artifacts such as ships; some constructed water features (will they more often be grouped with water features or with artifacts?); some terms for kinds of wetlands; and some additional land cover types or biomes, such as tundra, prairie, and savanna. Eventually we also wish to test similar sets of terms for other languages.

Another extension of this work would be to code the taxonomy in an ontology-coding framework such as Protegé. This will require the introduction of, and naming of, internal nodes for the taxonomy, ideally from an established Upper-Level Ontology such as DOLCE or BFO.

5. Acknowledgements

Comments from Gaurav Sinha and from an anonymous reviewer were useful and are appreciated.

6. References

Borgo, S., and Vieu, L., 2009. Artefacts in Formal Ontology. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 23, 3-21.

Egenhofer, M. J., and Mark, D. M., 1995. Naive Geography. In Frank, A. U. and Kuhn, W., (Eds.), *COSIT 1995*, Berlin: Springer, pp. 1-15.

Janowicz, K., Schade, S., Bröring, A., Keßler, C. Patrick Maué, and Christoph Stasch, 2010. Semantic enablement for spatial data infrastructures. *Transactions in GIS*, 14(2), 111-129.

Klippel, A, Wallgrün, J O, Yang, J, Mason, J S, Kim, E-K, Mark, D M, 2013, Fundamental cognitive concepts of space (and time): Using crosslinguistic, crowdsourced data to cognitively calibrate modes of overlap. In Tenbrink, Stell, Galton, Wood (Eds.), *COSIT 2013*. Berlin: Springer, pp. 377–396.

Smith, B., and Mark, D. M., 2001, Geographic categories: An ontological investigation. *International Journal of Geographical Information Science*, 15 (7), 591-612.

Wallgrün, J O, Yang, J, Klippel, A, Dylla, F, 2012, Investigations into the cognitive conceptualization and similarity assessment of spatial scenes. In Xiao, Kwan, Goodchild, Shekhar (eds.) *Geographic Information Science-7th International Conference, GIScience 2012*, pp. 212 – 225.

Wallgrün, J O, Klippel, A, & Mark, D M, A new approach to cluster validation in human studies on (geo)spatial concepts. *GIScience 2014 (extended abstracts)*.