MARK SCHILDHAUER
National Center for Ecological Analysis and Synthesis (NCEAS)
University of California, Santa Barbara
Email: schild@nceas.ucsb.edu

Mark Schildhauer is Director of Computing at the National Center for Ecological Analysis and Synthesis, NCEAS (http://nceas.ucsb.edu) at UCSB, where he oversees NCEAS’ cyberinfrastructure and technical staff, focusing on meeting the scientific computing needs of NCEAS’ associated projects and researchers.

Schildhauer’s technology research interests are in conceptual data modeling, the semantic web, scientific workflows, and Open Science—all in the context of facilitating integrative environmental and conservation science. Mark is currently involved in the NSF’s DataONE (http://dataone.org) Provenance and Measurement Semantics Working Group, and the NSF’s EarthCube GeoLink project (https://earthcube.org/group/geolink), which are advancing the use of techniques in Linked Open Data and the Semantic Web to describe and confederate disparate earth science data resources.

Schildhauer has a Ph.D. in Ecology, Evolutionary, and Marine Biology from the University of California, Santa Barbara, and an A.B. in Biology from Harvard.

Perspective

 How can spatially mediated discovery provide single-point access to research data, across distributed repositories and catalogs?

Access to spatial data is increasingly important to researchers in ecology and the environmental sciences, as the focus shifts beyond solely the experimentally-controlled “plot” scale to the landscape if not global scale of investigations. While the volume and variety of relevant spatial data for these investigations continues to rapidly grow, finding and harmonizing these data for synthetic analyses has been and will increasingly become more problematic. We believe this is due to several things. First, spatial data are broadly distributed—often housed in incompatible “siloes” due to idiosyncratic modeling of the “same” concepts, and these siloes not intercommunicating due to mismatches in terminology and schema. Knowing which of potentially numerous spatial data portals to visit, and then learning how to find and access the data one needs, becomes a time-consuming chore. Then, one finds that the data are documented with a variety of (often) sparse metadata, adhering to multiple potential standards or non-standards. There follows a conceptually and often computationally challenging and tedious process of transforming and integrating the data for one’s particular analyses. The difficulty of data integration is compounded by variation in the grain and precision of the data, as well as errors or omissions in their creation (i.e. variation in data quality).

How can we resolve some of these issues? We believe the development and more effective promotion of open-source, community-vetted, standardized models and vocabularies for describing the essential aspects of spatial phenomena and data—along with free, accessible tools and frameworks to link these to existing spatial data repositories—can alleviate some of these problems,
and represent a significant opportunity for improving this situation. The OGC is a good example of how these types of efforts can be productive, but there is danger of over-engineering solutions such that they become impractical for implementation, except for the most well-resourced participants. That is, highly complex metadata or other schemas and services can hinder uptake and participation. This is not ideal—both production and consumption of spatial data should ideally be democratized to the broadest possible population of consumers AND producers. In addition, any solutions should be fully interoperable with Web technologies—that is, involving globally unique, web-resolvable and dereferenceable identifiers for objects at grains of interest relevant for researchers, that are using semantically-defined terms curated by community-authorized consortia, and represented using W3-sanctioned languages, such as RDF/OWL.

- **How can the discovery of research objects in general be spatially supported?**

The advent of increasingly miniaturized and affordable GPS sensors, along with other sensor technologies, is making it feasible to provide highly accurate geospatial information along with any field or remote-sensed observations and measurements. In the case of ecology, biodiversity, conservation, and associated environmental sciences, however, while “raw” geospatial information can enable understanding as to the specific location, “distance-between,” trajectory, areal extent, etc. of the observations—it is the opportunity to **correlate a multitude of other measurements/observations through spatiotemporal coincidence** that will also be extremely useful. For example, a query such as: “Where will drought most severely impact agricultural productivity” recommends spatial resolution to advance both general understanding and planning purposes—but fuzzy notions such as “drought” and “agricultural productivity” will need to be defined, as well as linked to locations in order for such queries to be resolvable. The spatial boundaries on the locations to which these thematic terms are linked may also be fuzzy or probabilistic: representing this is a research challenge.

*Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line.* (Mandelbrot)

The fractal geometry of nature, coupled with “fuzziness” can be particularly problematic when referring to **named** spatial phenomena. Is “Santa Barbara” part of “Southern California”? While spatial ontologies MUST address issues of membership, containment, adjacency, disjointedness, etc. (and medical anatomy ontologies provide some good precedents for how to do this), and do so in a consistent, unambiguous, web-accessible way—natural language allows for many variations on how terms can link to the concepts they represent. Words that are polysemes, or have homonyms, synonyms, or plesionyms create challenges relative to appropriate or useful query resolution. Named places in particular often are polysemous, especially if not bounded by fiat. Temporal instability of spatial structure and configurations add another element of complexity to recognize and deal with.

These complications further recommend resolving the first point above, to enable spatial information to more effectively become an important “Key” linking thematically described data to other data of potential interest.
• How can spatial discovery be applied to topic spaces, not just geographic ones?
I am not clear here if “topic space” is being used in a narrow “Publishers” sense of being able to group thematically related sets of publications. But whether the connotations of “topic spaces” here are narrowly circumscribed, or intended to be interpreted more broadly, we’d recommend that the approach for developing “topic spaces” and assigning digital objects as members to one or more of these “topic spaces,” conform to emerging standards for the W3C semantic web approaches, specifically by being represented in formal RDF/OWL ontologies. This logic-based language enables us to define Classes that can have Instances, and Instances are members of Classes. Classes can also contain SubClasses, and Instances can be richly related to one another through Object Properties assigned to the Classes. The logical expressivity of OWL allows one to build a rich and W3C standards-interoperable specification of how objects and their digital representations might “belong” to various “topic spaces” (although again, I’m not familiar with the topic space approach).

Finally, at the meta-level, I still find the conceptual model of accumulating “thematic-defined, spatially-registered layers” to be both intuitive and useful.