Mapping Natural History Museum Data with GIS:
Conservation Use, Retrospective Georeferencing, and Collaborative Efforts

Michelle S. Koo[1] and Elizabeth J. Proctor[1,2]

[1]California Academy of Sciences, San Francisco, CA; [2]San Francisco State University

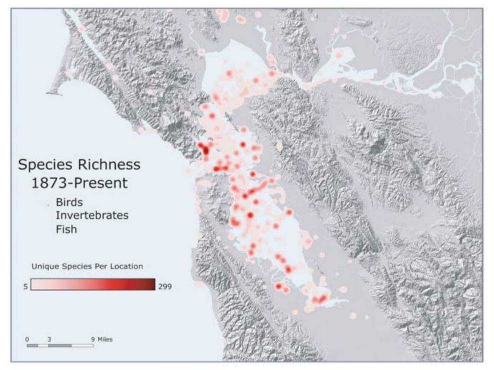
Abstract

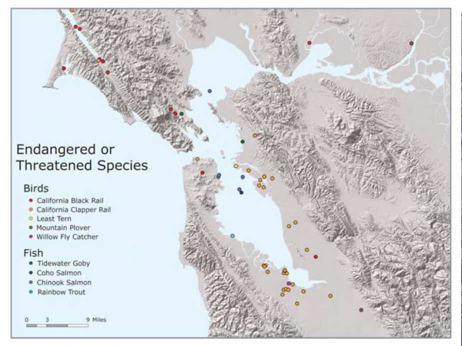
Conservation authority and decisions rely on solid natural history science. *What species exist? What kind of habitat is involved? Where are they occurring? What is happening over time?* GIS is the technology that facilitates sophisticated answers to these questions, and acquiring natural history information into quantitative spatial terms is recognized as a priority among the repositories of such data, the natural history museums. The task of retrospective georeferencing invites an opportunity to consider conceptual issues of how location is interpreted, from written description to spatial expression. Variations in interpreting localities of specimens exist due to the human-mediated process. Collaborative efforts to georeference common taxonomic collections "pool" their localities to maximize efficiency, use geographic "expertise" of individual institutions, establish repeatable protocols, and reduce variation in interpreting locality. The online dissemination of georeferenced data by collaborating institutions will propel geospatial biodiversity studies of species and their natural history.

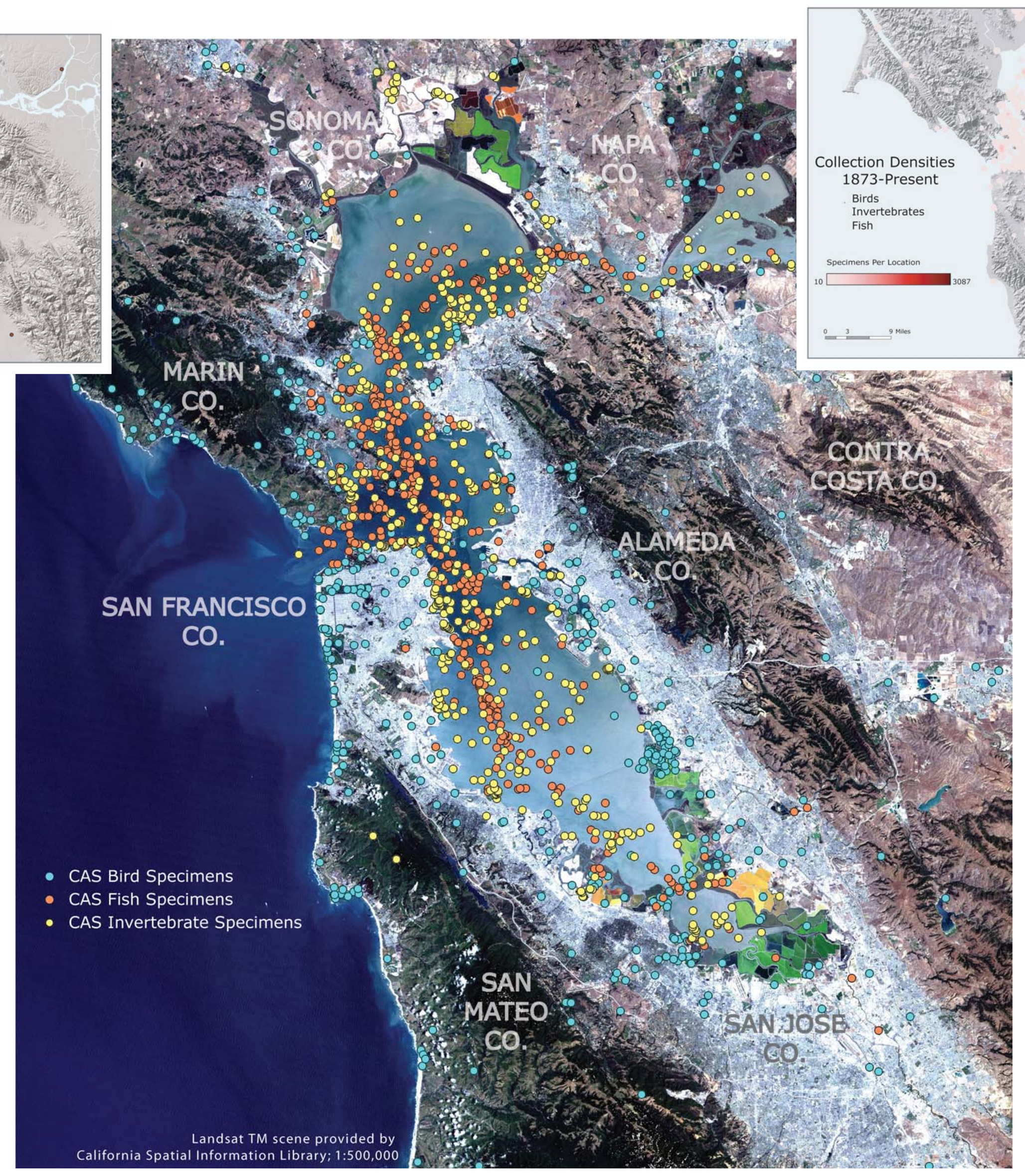
Informed Conservation Decisions Start Here

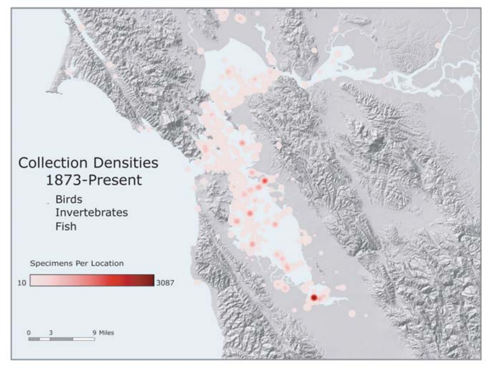
Natural history specimens and their associated data are integral to biodiversity research that informs conservation policies and decisions. Collections of floral and faunal specimens represent historic baseline data from which species richness, endemism, extinction rates, former ranges, etc. can be determined.

Here, California Academy of Sciences' (CAS) collections of coastal bird, fish, and invertebrate species from San Francisco Bay spanning 1873 to 2003 are shown. At right, density grids summarize collection holdings.





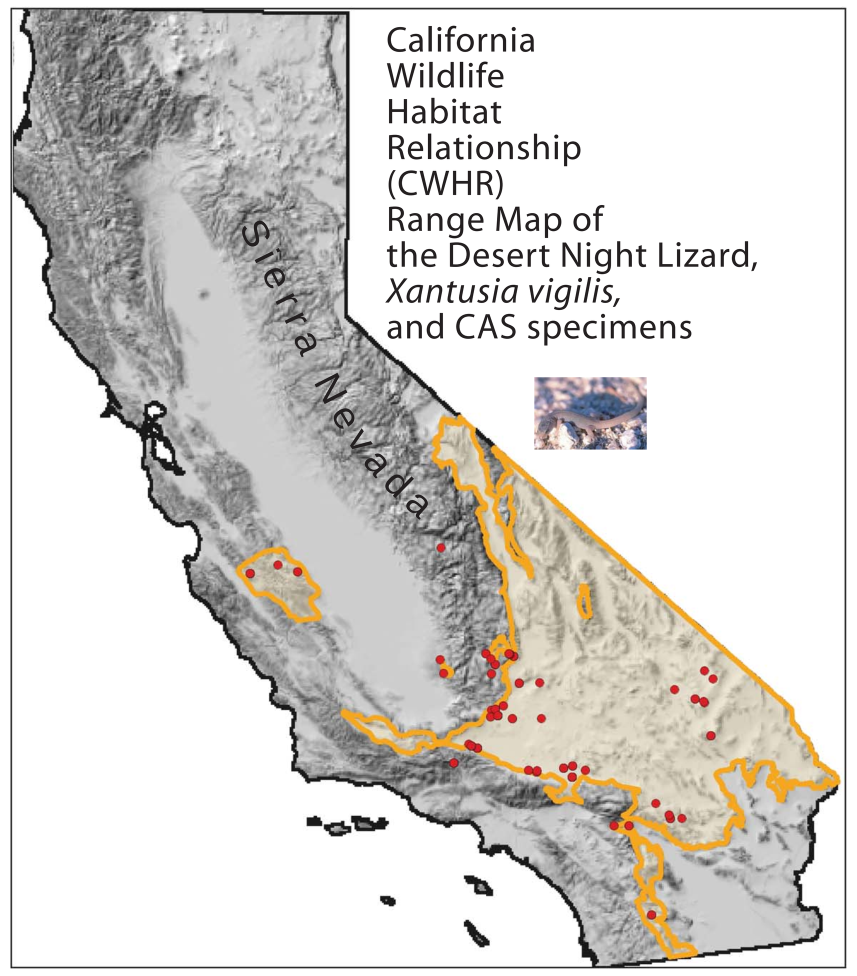




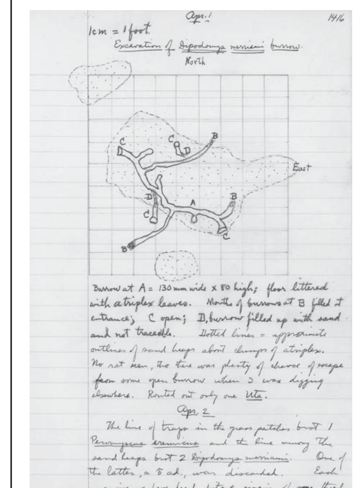
Example of Immediate Use

Range maps of species are based, in part, on museum specimens as well as generalized habitat preferences, vegetation maps and elevation models. Museum specimens also verify and constantly refine the known ranges.

Red circles on the state map show historic and recent specimen localities; a 2002 locality falls outside the presumed range within the Sierra Nevada, provoking questions of the species' biogeography.

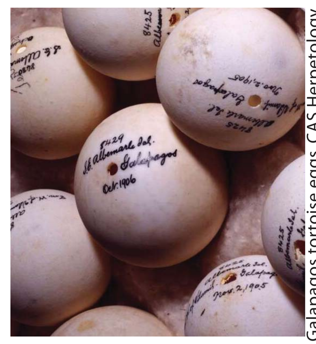


The Basics of Museum Data



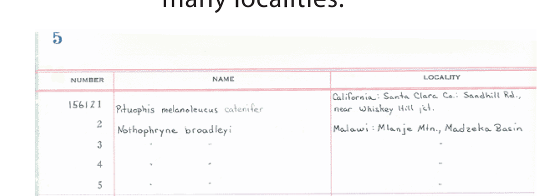
J. Grinnell 1917 field note page, courtesy Karen Klitz, Museum of Vertebrate Zoology, U.C. Berkeley.

Collecting and cataloging animal and plant specimens are traditions that date back hundreds of years. Housed in natural history museums and research institutions worldwide, millions of preserved animal and plant specimens along with associated data such as the collection date and location represent defensible occurrence data that can reveal trends in animal and plant distribution and biodiversity over time. The traditional use for these collections is to support studies in systematic biology, but increasingly, they are being used for spatial analyses.



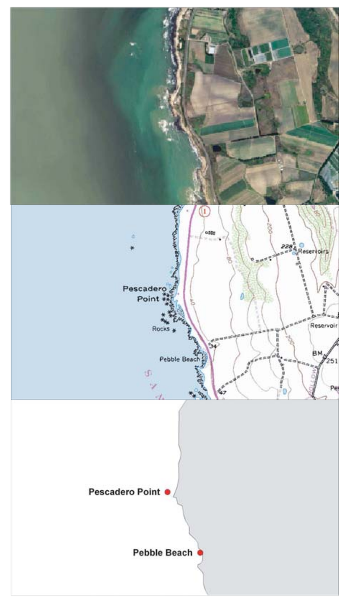
Galapagos tortoise eggs, CAS Herpetology

Geocoding natural history specimen localities is fundamentally different from the common sense of 'geocoding' as it does not provide location by street map reference. The table at right shows typical locality descriptions, from place or feature names to offsets from named places, and the typical vagueness inherent in many localities.



RECORD	LOCALITY
1	Millbrae
2	Belmont area
3	San Bruno Mountain
4	Pescadero Point
5	San Francisquito Creek
6	head of Lobitos Creek
7	1 mi S Moss Beach
8	0.33 mi SE Sharp Park Golf Course
9	tower S end San Andreas Lake
10	N of La Honda
11	2 mi E of State Hwy 35, along Kings Mt Rd

Retrospective Georeferencing and Interpreting Localities

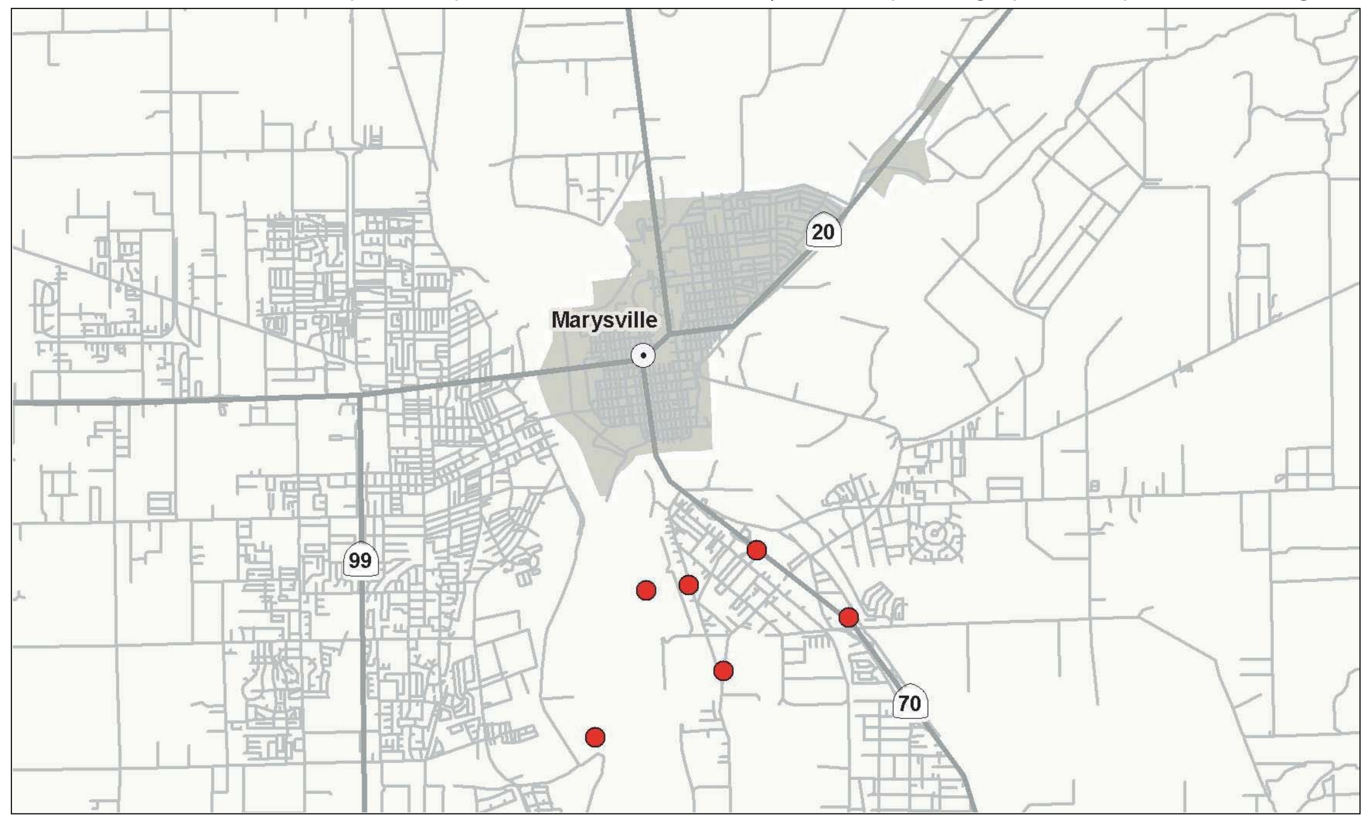


Any computer application using specimen data is limited by the availability of proper spatial location data as input. Most historical locality data are composed of textual place descriptions rather than geodetic coordinates. In other words, "1 mi S of La Honda" describes where a specimen was collected rather than a coordinate pair of longitude and latitude. No specimen locality can be mapped with GIS until this textual description is "translated" into coordinates.

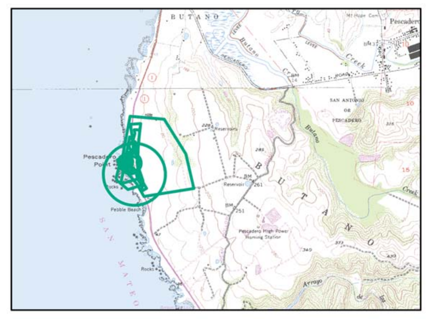
Many institutions recognize this need by going through their historical specimen localities and assigning coordinates, a task called "retrospective georeferencing." Different methods are currently being used, but in most cases georeferencers refer to a digital basemap such as a USGS topo quad and then draw a point or polygon to represent each textual place description. The resulting coordinate pair or georeferenced shape is then associated with the specimen's record in the collection database.

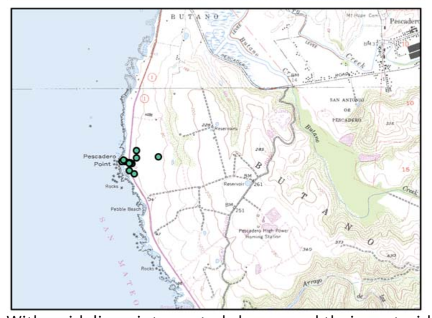
Although automated computer techniques are being explored, most current retrospective georeferencing efforts are human-mediated and thus involve some amount of subjectivity and interpretation.

Example: The locality "2 mi S Marysville" can be interpreted several different ways: it could be two miles measured along a road leading south out of Marysville, or two miles directly south "as the crow flies." Furthermore, two miles could be measured starting from the center of town or from its southernmost edge. Depending on the year the locality was recorded, the city limits of Marysville themselves may have changed. Red dots indicate six minimum possible placements for "2 mi S Marysville" depending upon the spatial reasoning used.



Variations in Interpretation



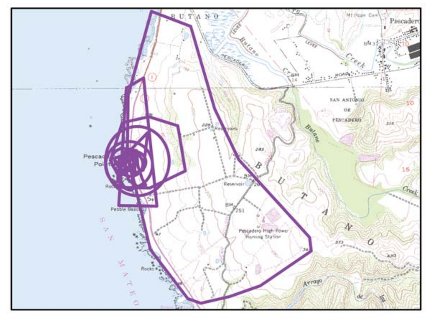


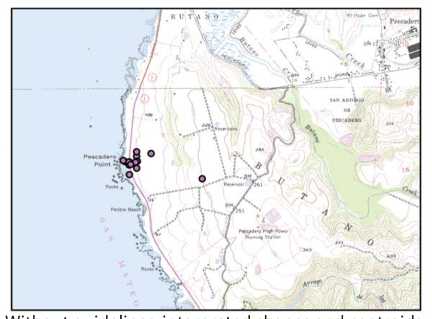
With guidelines, interpreted shapes and their centroids for the locality "Pescadero Point" (Proctor 2004).

The concept of "place" is not fixed and subject to interpretation. Given what is known about human spatial reasoning and perception, it is reasonable to expect individuals to interpret the same textual locality phrase different ways. If the locality is vague, the variation in interpretation among individuals will be even greater.

As retrospectively-georeferenced coordinates begin to populate biological databases worldwide, this variation is of concern because having different coordinates purported to represent the same location is poor data management and introduces uncertainty to subsequent spatial analyses.

In practice, some institutions attempt to reduce this variation by providing explicit instructions on how to interpret problematic localities. The hope is that guidelines will reduce the frequency of subjective judgment calls about vague places and increase homogeneity among georeferenced localities.





Without guidelines, interpreted shapes and their centroids for the locality "Pescadero Point" (Proctor 2004).

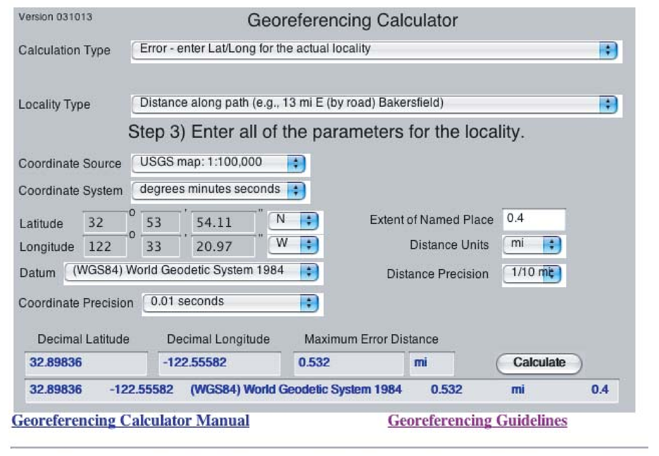
A recent thesis study (Proctor 2004) measured the variation among shapes drawn by test subjects interpreting a set of 11 textual localities, and to what extent instructions reduced that variation. The results indicate that the guidelines provided in the study did not make a statistically significant difference between the two groups in shape location or shape size. One interpretation of the results is that people tend to follow reasonably similar logic when interpreting localities, with or without guidelines. The guidelines appear to have the effect of refining shape sizes and placements in small but noticeable ways and have little effect on interpretation of the vaguest localities.

Methods and Metadata

With interpretive variation and vagueness in historical specimen location, documenting the quality of georeferenced data is critical. By far, the most frequently adopted method has been to categorize locality accuracy into ranked or ordered classes (e.g., a series of codes indicating precision level). This type of coding has been widely criticized as fundamentally arbitrary, inappropriately presumptuous about user requirements, and unlikely to be universally accepted.

By contrast, recent collaborative georeferencing projects utilize a methodology for recording locality imprecision quantitatively. HerpNET and MaNIS project participants assign each locality a pair of coordinates and a corresponding estimated error distance value. This maximum error distance is interpreted as a radius around the coordinate point, calculated quantitatively from identified sources of "uncertainties" in the locality description. The error distance values can be used as a measure of locality vagueness or confidence, allowing subsequent users to exclude locality data that are too vague for their particular application.

A locality's ultimate uncertainty distance value is calculated based on the relationship and interactions of all its uncertainties. Example: "6 km E (by road) of Bakersfield," the guidelines identify four sources of uncertainty (the extent of Bakersfield, unknown datum, distance imprecision, and map scale) and provide instructions on how these interact and how to calculate a value based on these particular interactions (in this case, 4.051 km). Some interactions are additive and linear, some are nonlinear.

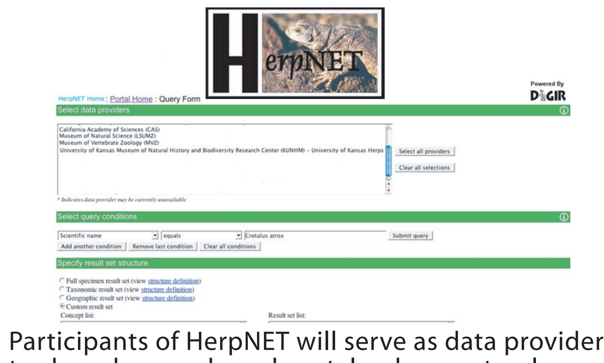


The 'Georeferencing Calculator,' an online web form, has been developed to facilitate calculation of complex uncertainty distance values.

Collaborative Efforts, Sharing Data

The National Science Foundation has funded collaborative efforts to georeference historic localities for collections of natural history specimens, such as mammals (MaNIS), fish (FishNET), amphibians and reptiles (HerpNET), and potentially birds (OrNIS) that can be distributed over networks of data providers and data portals. Collaborating allows coordinating geographic expertise and resources, and more importantly, establishes repeatable procedures. Because of the unique nature of specimen localities, new online tools have been developed to facilitate georeferencing and calculate error estimates.

The products of these collaborations will seed a GIS of natural history specimens that can be used in spatial models of predicted species distribution, verification of species occurrence, or shifts in faunal assemblages through time, essentially the science to inform conservation.



Participants of HerpNET will serve as data providers to shared, query-based portals where natural history data, including georeferenced locality, can be accessed.

Websites to projects:

Reducing Variation in Georeferenced Locality Descriptions
Elizabeth J. Proctor 2004 — www.calacademy.org/research/informatics/georef/ProctorEJ_ThesisAbstract.html

http://elib.cs.berkeley.edu/manis/

http://www.herpnet.org/

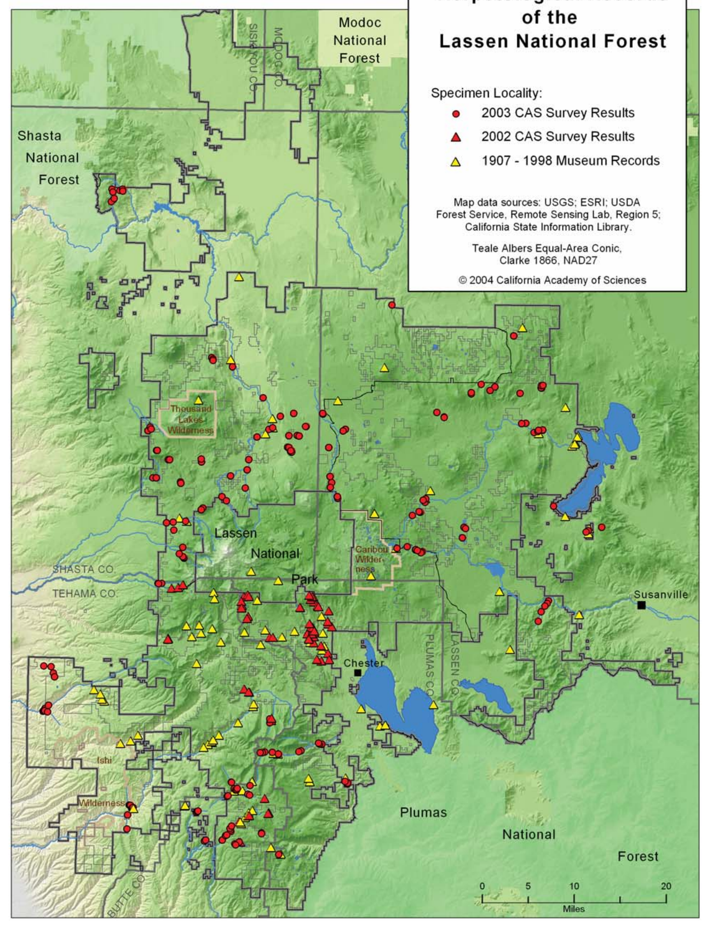
European Natural History Specimen Information Network — http://www.nhm.ac.uk/science/rco/enhsin/introduction.html

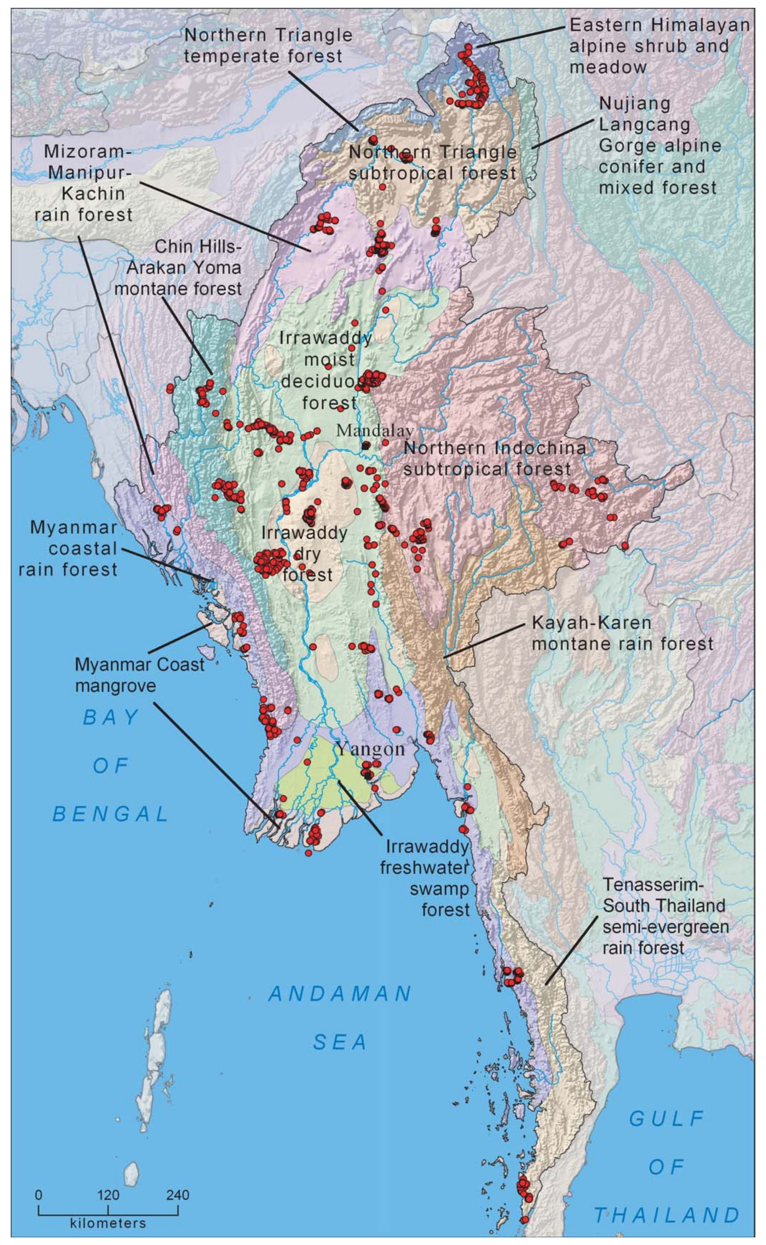
Ongoing Surveys

While museum resources are valued for their historic breadth, the ongoing science of surveys continues.

Historic museum data informed World Wildlife Fund's project to delineate global terrestrial ecoregions. At right, collection sites (red circles) for amphibians and reptiles from the recent CAS Myanmar Herpetological Survey Project are compared to ecoregions.

Below, historic sites (yellow) help guide survey priorities (red) in the Lassen National Forest, California, in an ongoing CAS partnership with the Forest Service to help their wildlife biologists determine species occurrence, especially listed amphibians and reptiles.





August 2004

Contact: mkoo@calacademy.org

Image from The Natural History Museum, London