Bone Rooms, Bird Bodies, and Biodiversity Informatics

Jonathan W. Lowe

Where in the world does the same grizzly bear perpetually whack the same leaping salmon — both creatures frozen in shared savagery — while children whisper and point only inches away? Some people believe that museums contain only musty air, stuffy docents, and pure boredom. However, tucked away behind a mysterious door marked "Museum Staff Only" is a dynamic and ever-growing resource few of us are lucky enough to see in person: the museum collection itself. Whether you imagine graybeards stirring up dust as they pin shiny beetles into tiny boxes or a sparkling modern facility, every museum's beating heart is its hidden collection of specimens and associated library of descriptive notebooks. These collections are anything but boring, and many are now online.

Dust or no dust, it also may be difficult to guess how 100-year-old bird bodies could have any relevance to the geospatial industry or our lives in general. One visit to the consortium of Berkeley Natural History Museums (BNHM, http://bnhm.berkeley.museum), however, reveals that collections (worldwide) are data storehouses of tremendous relevance to researchers in such disciplines as biology, geomorphology, ecology, and climatology. A natural history museum, in particular, is not just a warehouse of dead creatures, but a spatio–temporal census of flora and fauna. Need to search 100 years of specimens for mammals collected in Colorado, sorted by genetic signature and mapped by evolving distribution? Museum curators teaming with in-house geospatial experts are enabling just such analyses by developing spatio–temporal specimen catalogs, taxonomic protocols, and online spatial visualization tools capable of geocoding even "fuzzy" data from historic textual references.

Many museums are digitizing spatio–temporal data about their collections and making them available online, providing a valuable research tool for increasing knowledge about our world.

Chasing Critters

What our industry typically calls geospatial data — points, lines, polygons, raster grids, and so forth — are either ink on paper or electronic zeros and ones. Usually, when we venture into the field to collect vector and raster data, we don't really bring anything tangible home. The street centerlines we digitize merely represent the streets — we capture the bits and bytes, but leave the asphalt where it is.

A stuffed albatross (above) presides over the MVZ specimen collection, housed in dozens of metal shelves filled with hundreds of trays of preserved creatures.

Net Results columnist Jonathan W. Lowe covers the role of emerging technologies in the exchange of spatial information. Lowe is the owner of Local Knowledge Consulting (Berkeley, California), where he designs and implements spatial Web sites. He can be contacted at info@gisswebsite.com.

Museum data are, quite literally, a different animal. Museum collectors note when and where they found the creatures they were seeking, but also often bring some critters home with them for further study and preservation. It's both the collectors' written records and the actual bodies, bones, and skins that fill a museum collection. (In comparison, a GIS lab seems a bit empty and sterile — just a few posters and some humming machines.)

In the past, field biologists collected specimens with shotguns, leg-hold traps, or snap traps. Modern collectors are more likely to capture and release most of the animals they discover, keeping only enough specimens for positive identification and later reference (particularly when sampling small mammals, amphibians, or reptiles). Researchers may trap in the morning and then remove the animals' skins and stuff them with cotton in the afternoon. Back at the museum, they tag the skinned bodies and drop them in a tank of flesh-eating beetles that leave only bones behind. To capture the DNA, collectors save small slices of the animals' livers in vials of alcohol.

Some research calls for data about whole groups of animals rather than individuals, such as with bird population studies. In this case, naturalists simply observe and count, bringing home only photos and notebooks. The contents of the notebooks are data, of course, but the notebooks themselves may also become historic specimens over time. Berkeley's Museum of Vertebrate Zoology (MVZ), a member of the BNHM consortium, for instance, has journals from such collectors as Joseph Grinnell and Aldo Leopold (author of *Sand County Almanac*) that are as worthy of preservation as the specimen collections they describe (see Figure 1). Grinnell's highly detailed field notes established a system in the early 1900s that continues to this day at many museums. Specifically, Grinnell attached tabular data to each specimen using a consistent organizational template — in other words, he pioneered a metadata standard for museum collections.

Part of that standard includes specific spatio–temporal metadata. For instance, Grinnell and his colleague, Tracey Storer, conducted a survey of birds, mammals, reptiles, and amphibians from California's Central Valley through Yosemite Valley to Mono Lake between 1914 and 1920. They followed a transect — one line cutting across many different ecosystems — that they gradually navigated during six years of field work. As they captured and observed the creatures, the researchers noted both location along the transect and date of each capture. Today, nearly 100 years later, new curators at the same institution, Berkeley's MVZ, are following that same transect to detect changes in species abundance and distribution.

Such long-term comparison studies, however, raise issues about incompatible data formats. It's safe to assume that any modern field-collected data, even if not captured digi-

Glossary

BNHM: Berkeley Natural History Museums

DiGIR: Distributed Generic Information Retrieval

GARP: Genetic Algorithm for Rule-set Production

KUNHM: University of Kansas Natural History Museum and Biodiversity Research Center

MaPSTeDI: Mountains and Plains Spatio–Temporal Database Informatics Initiative

MVZ: Museum of Vertebrate Zoology

NSF: National Science Foundation

XML: Extensible Markup Language

A tray of colorful South American bird specimens tagged with collection metadata.

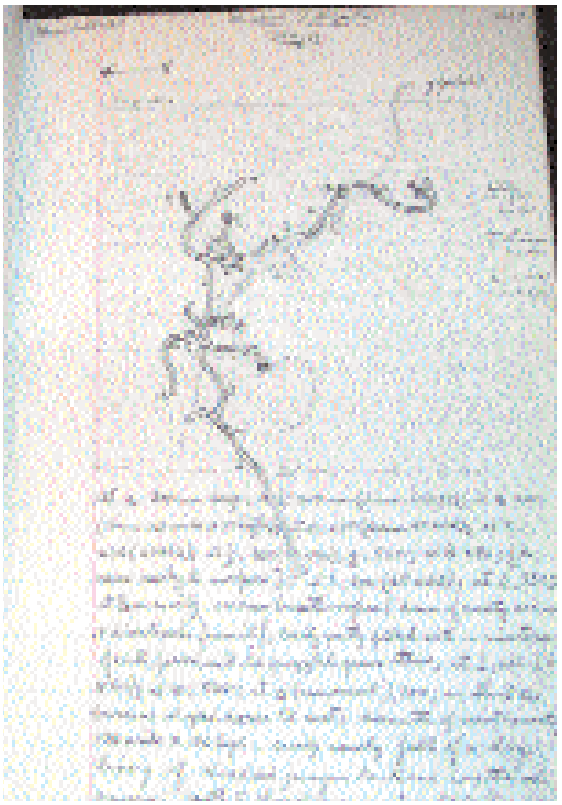

FIGURE 1 A page from one of Joseph Grinnell's 1918 notebooks shows his penciled map of gopher burrows in Siskiyou County, California. Grinnell's notebooks, which are now being scanned to create to a queryable Internet-based database, are rich in (spatio–temporal) textual references to historical ecologic conditions.

tally, will ultimately be converted to digital format. Data collected before computers even existed, such as that in Grinnell's and Leopold's notebooks, are also valuable when making temporal comparisons with parallel studies today. Consequently, starting in 2003, the National Science Foundation (NSF) awarded MVZ a grant to scan Grinnell and others' 13,000 pages of field notes and 2,000 photos to a queryable Internet-based database. The notebooks' text will become searchable by specimen catalog numbers, names of collectors, scientific names, common names, places, and dates.

Supporting the digitization effort, a program called BioGeomancer (www.biogeomancer.org) can accept even "fuzzy" textual spatial references such as "seven miles west of Davis," and automatically return a point location. BioGeomancer is the result of a partnership between the University of Kansas Natural History Museum and Biodiversity Research Center (KUNHM, http://nhm.ku.edu), Brazil's Reference Center on Environmental Information (www.cria.org.br), Yale University (www.yale.edu), and MVZ (www.mip.berkeley.edu/mvz). BioGeomancer's founders have named the service's capability "geoparsing," and like any well-designed Web service, it provides just that single function. The Web site's interface is consequently (deceptively) simple, offering users four text entry fields beginning with country, stepping down in scale through state and county, and ending with locality. The service will format geoparsed results as hypertext markup language, extensible markup language (XML), or a graphic map.

BioGeomancer matters to collectors, curators, and users of natural history specimens, because it extends the gazetteer concept to handle the grammar that biologists in the field commonly use to describe locations. Basic gazetteers convert place names of, say, cities or monuments into points. BioGeomancer's enhancement to the gazetteer concept is that it parses not just single place names but whole phrases, including locations at some distance and cardinal direction from a nearby city or monument. When hunting for specimens, collectors are seldom actually in the cities that gazetteers reference, but often refer to their position in relation to a nearby city or distant mountain peak.

A parser (in this context) means an algorithm that recognizes the syntactic structure or grammar of a text phrase, and so can plug the phrase's individual words into a data model, then perform some calculation on that data, such as geocoding. BioGeomancer is clever enough to deconstruct the jumble of possible grammars that may frame a spatial phrase. For example, BioGeomancer can successfully parse each of these three phrases despite their different structures: "2.4 km WNW of Pandemonium," "Springfield, 22 miles E," and "Springfield, 0.5 mi. E of Pandemonium." Once the parser recognizes how the words in a phrase reference a location, the service returns that location's latitude and longitude. BioGeomancer also supports batch geoparsing through several application programming interfaces including a simple object access protocol/XML interface.

Multidimensional Categorization

BioGeomancer's geoparsing of textual references solves part of the digital conversion problem, but what's the best approach for museums hoping to digitize metadata for the tens of thousands of physical specimens in their collections? Each specimen may have taxonomic, genetic, and spatio–temporal metadata elements. And researchers may want to search for specimens based on any or all of these criteria. Given the disarray and worldwide distribution of museum collections, standards become critically important. For example, the museums of the Russian Academy of Sciences in St. Petersburg contain significant numbers of California specimens from the mid-1800s — when Russian otter-trapping expeditions visited North America's West Coast. Worldwide researchers of California species need to be able to search for records of, say, *Ursus arctos* (Grizzly Bear) and receive a list of specimens in both California and Russian museums.

Because museums capture different details about their specimens using different metadata structures, search engines can rely on only a handful of common metadata elements in worldwide collections.

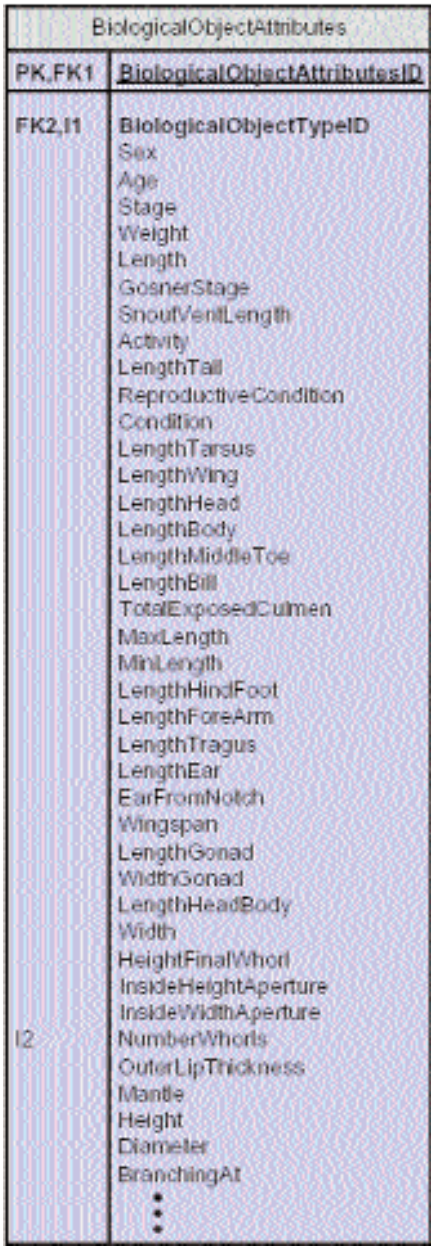

FIGURE 2 Specify's logical data model gets very specific about measurements of common body parts, as illustrated by this subtable of bird specimen metadata.

However, once a researcher locates an individual specimen, she is likely to want much more detail. To understand the depth of detail available for some collections, consider the data model of a system called Specify (www.specifysoftware.org). The model includes metadata slots for a specimen's taxonomy, physical characteristics, and even the methods by which it is preserved, among others (see Figure 2).

In theory, once museums adopt a common data model for their specimen meta-

data, they can then leverage a common query format to search all archives. Although it involves change-management issues, the Darwin Core version 2 model is currently the most widespread schema for specimen metadata.

The Network Is the Museum

Even if collections keep standard metadata about their specimens and notebooks, researchers still need a protocol for querying those databases (or even finding them in the first place). As described in a briefing by the Task Group on Access to Biological Collection Data (www.bgbm.org/TDWG/CODATA),

> The world's collection databases represent myriad database and access technologies. Many of these are primitive, hard to use, platform specific, and scale poorly. Furthermore, almost all of the existing systems are incompatible with each other. Rather than encourage database providers to take on the burden of redesigning/rebuilding existing databases, the software architecture track defined a system of 'gateways' to wrapper around existing databases and 'portals', which would access the gateways . . . [leading to] decoupling of portals and providers. With standard data provider software and capabilities, any organization with special skills in data integration (perhaps beyond biological collection data, such as GIS data) and designing easy-to-use interfaces should be capable of establishing a portal to collection data.

The Task Group has consequently developed a protocol (that is, an agreed-upon way of exchanging information) for sending queries to biological collections' databases. The protocol is DiGIR, the Distributed Generic Information Retrieval protocol (www.digir.net), and it is a more flexible alternative to earlier efforts' reliance on the Z39.50 protocol. DiGIR's architects designed the protocol to be both technically simple (relying on hypertext transfer protocol and XML) and flexible enough to embrace a growing population of museum participants regardless of their existing database technology.

For the record, cobbling together distributed databases is no small computing endeavor. Consider, for instance, the numbers involved in a database unification project at KUNHM that employed the Species Analyst (http://speciesanalyst.net) approach. When researchers pooled just 12 digital fish collections, they gained access to 20 million fish specimens. And because most large collections are only partially digitized, the data volumes can only grow. As John Deck, BNHM Informatics coordinator, notes, "Our work of digitizing specimen labels has just begun. We only have 12 percent of our 15 million specimens done. We've got the mammals; the majority remaining are insects."

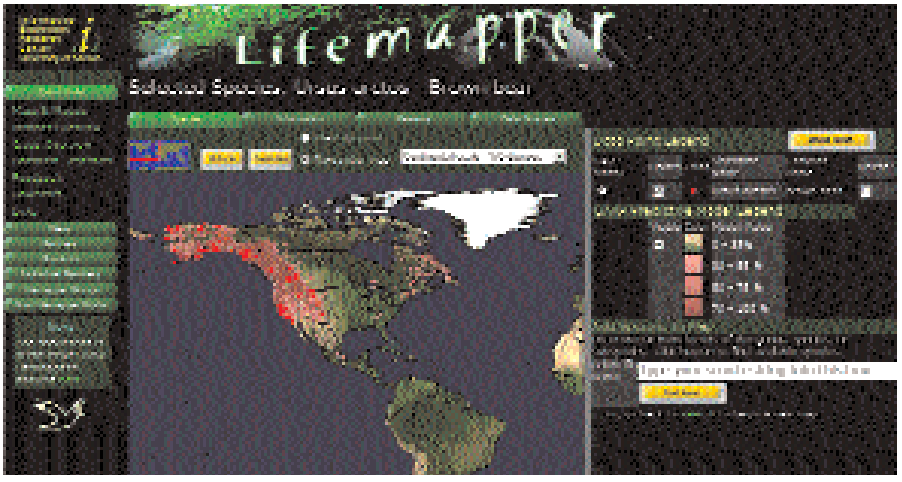

FIGURE 3 Lifemapper displays the results of a GARP model that extends point observations to predict the likelihood of species distribution using underlying ecology layers. This view maps Grizzly Bear data in North America.

Analyzing Species Distribution

Some museums that are ready to share their digitized collections use the Global Biodiversity Information Facility (www.gbif.net), a system for referencing distributed databases using such data models as Darwin Core in conjunction with the DiGIR protocol. But are there tools for manipulating the query results?

The growing availability of multi-institutional data has spurred other related projects, such as the Mountains and Plains Spatio–Temporal Database Informatics Initiative (MaPSTeDI, http://mapstedi.colorado.edu) and its online mapping interface, GeoMuse (http://mapstedi.colorado.edu/geomuse.html); LifeMapper (www.lifemapper.org) and its underlying analysis engine, Desktop Genetic Algorithm for Rule-set Production (DesktopGARP, www.lifemapper.org/desktopgarp); and standalone desktop products such as Diva (www.diva-gis.org). All these tools help researchers visualize and analyze collection queries online.

Lifemapper and DesktopGARP are NSF-funded projects for creating a comprehensive species distribution map archive. Lifemapper's online mapping tool plots specimen search results as points on a map (see Figure 3). It's also possible to drag a box on the map and get back all the specimen records for that location. Researchers use Lifemapper to plot every spatio–temporal point observation of a given species, then cross-reference the points with their underlying ecology (an overlay of elevation, rainfall, vegetation, ecological communities, and so forth) in order to extrapolate beyond the points to broader viable distributions of that species. In other words, Lifemapper can identify everywhere a species is capable of surviving. For instance, suppose collectors follow a single transect from the base to the top of a mountain, and observe squirrels along the transect between 3,000 and 4,000 feet of elevation. Unless the ecology of the mountain changes dramatically from one face to another, it's likely that squirrels live not only along just the transect line, but also in a band circling the whole mountain between 3,000 and 4,000 feet of elevation.

Another online species collection, MaPSteDI, contains biodiversity data exclusively for the southern and central

Rockies and adjacent plains. MaPSTeDI's online mapping tool, GeoMuse, is similar to Lifemapper in helping researchers graphically track changes in biodiversity. For example, searching for *Erigeron* (Daisy) specimens collected in Colorado between 1930 and 1950 returned several records and map locations (see Figures 4a and 4b). GeoMuse automatically displays the most appropriate background for each zoom level, such as U.S. Geological Survey (www.usgs.gov) digital ortho quadrangles and, at small scales, digital raster graphics (see Figure 5). Importantly, although museum data may seem to be an unexpected source of privacy issues, nonhuman species need protection too. GeoMuse does not provide geographic coordinates for sensitive collections such as paleontologic and endangered species (tomb robbers and poachers need not apply).

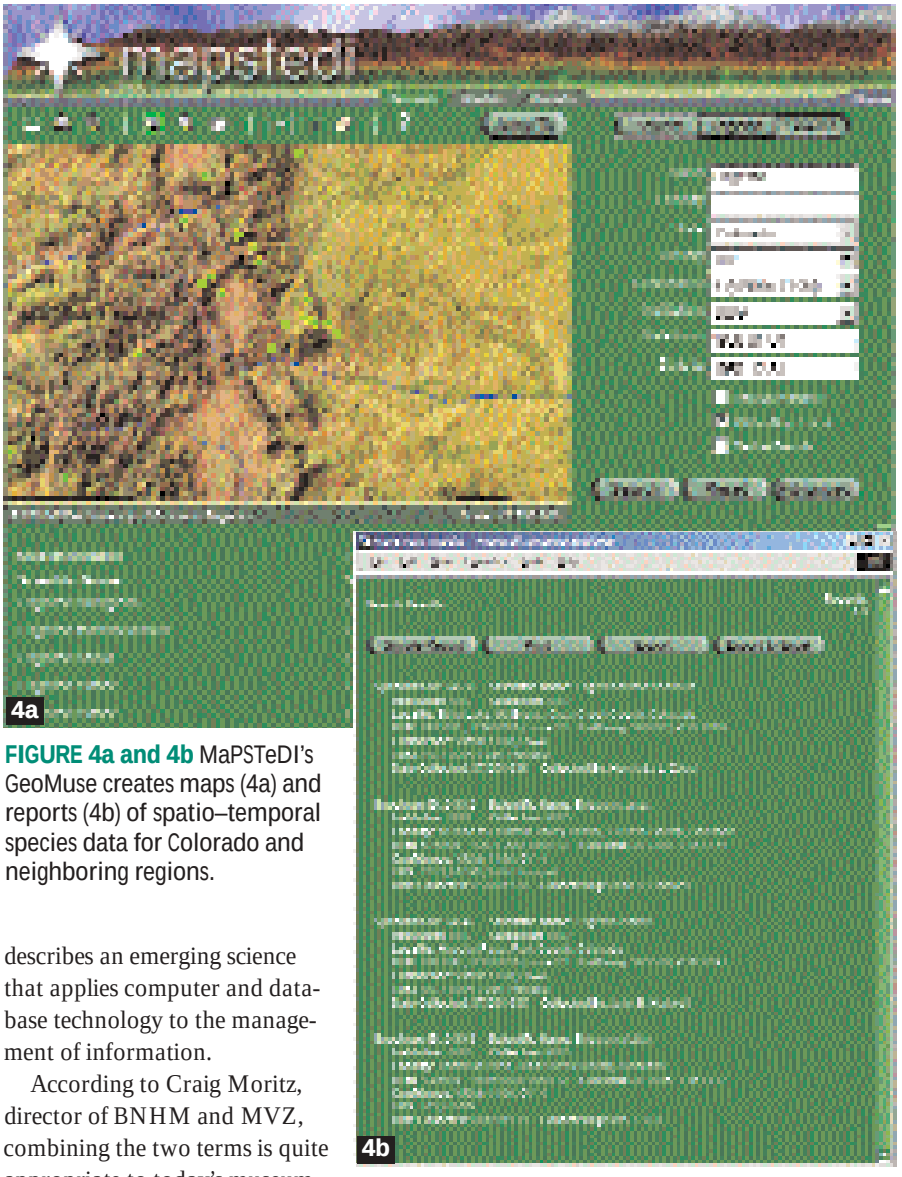

FIGURE 4a and 4b MaPSTeDI's GeoMuse creates maps (4a) and reports (4b) of spatio–temporal species data for Colorado and neighboring regions.

Biodiversity Informatics?

As demonstrated by the online tools discussed in this column and similar projects, such as the WhereWhy project (http://biodi.sdsc.edu/ww_home.html), the connection between observations and ecology supports several broad areas of research. By modeling changes to the ecology as a result of global climate change, researchers can predict impacts on terrestrial and marine biodiversity. Conservationists can find gaps in networks of conservation reserve systems that prevent species from moving freely through larger ranges. The field of biogeography can map evolution over time, using DNA signatures to isolate the changing distribution of a single species. Biologists can use the data for systematics and migration pattern research. Agriculture researchers can analyze insect collections to predict the timing and spread of pests. And anthropologists can follow the extinction of species from certain geographies to better understand and predict human habitation patterns.

As digital museums increasingly become resources for such research projects, they are transforming into warehouses for "biodiversity informatics" — a common term at online museum sites. Biodiversity is the relative abundance and variety of plant and animal species and ecosystems within particular habitats. Informatics describes an emerging science that applies computer and database technology to the management of information.

According to Craig Moritz, director of BNHM and MVZ, combining the two terms is quite appropriate to today's museum projects. He explained that "direct access to digital collections unites genome data, historical data, and even handwriting specialists, in exciting new collaborations."

Like so many other "informaticizing" institutions, museums are microcosms of the emerging trends in the computing and geospatial industries. Their efforts require standardization, use of modular and decoupled computing architectures, and increasingly broad cross-disciplinary access to formerly stovepiped datasets.

Moritz went on to add, however, that "Museum data is not perfect, so users should carefully validate the quality of digital databases, checking for variations in taxonomy over time and margins of error in geocoded points. Though we appear to the Internet user to be one enormous digital collection, not all participating museums share a common set of methods."

Taking a long-term view of online digital collections, Moritz observed that once the private backroom collection becomes publicly available online, the feedback loop between users and the curators that maintain the collections is much stronger. As users notify curators of inconsistencies, the quality of the collections improves rapidly.

Another result of widespread digitization and consequent overlay of all this disparate data is increased knowledge about our world. The Grinnell transect is a case in point. Upon comparing past and present collection results in Yosemite National Park's Merced Grove and Crane Flat

areas, researchers found that the golden-mantled ground squirrel had disappeared, moving to elevations 500 feet higher. Likewise, the piñon mouse appeared as high as 10,240 feet in upper Lyell Canyon and Glen Aulin; it was previously found only in the Eastern Sierra at elevations below 8,200 feet. Though it's too early to be certain, long-term patterns of animal migration to higher ground could signal global warming. Or they could be due to Yosemite's fire suppression policy, or to 1920s logging practices — it's still anyone's guess.

Whatever the reasons, simply recognizing the spatio–temporal changes more rapidly and accurately will not only increase our ecological understanding, but will guide future land stewardship. Who knows what other discoveries await us at this confluence of bone rooms, bird bodies, and biodiversity informatics?

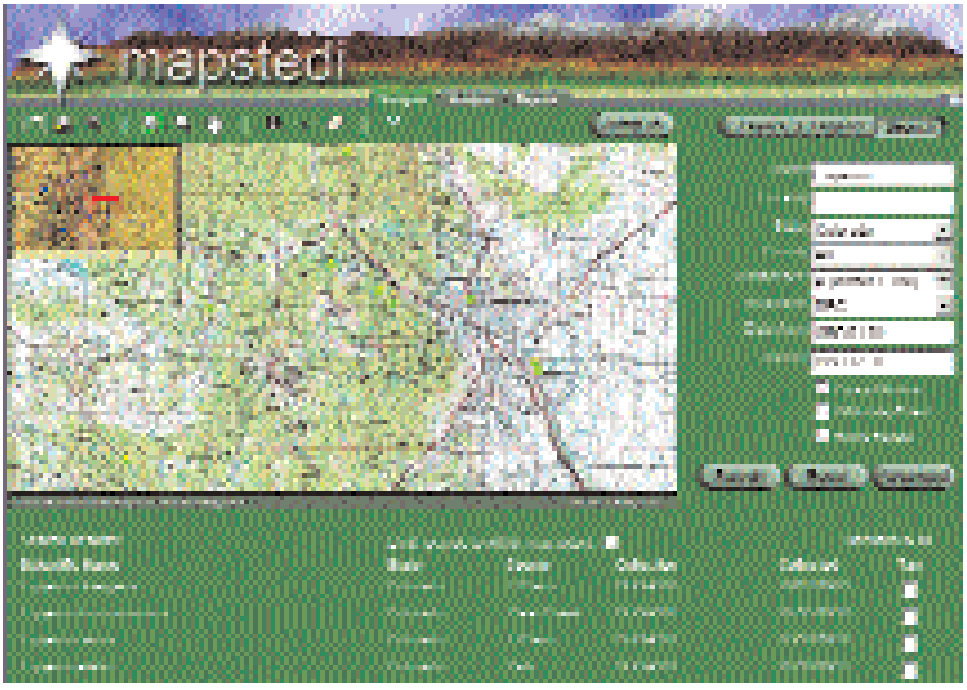

FIGURE 5 Zooming in on a GeoMuse map replaces the color-relief background with U.S. Geological Survey digital ortho quadrangles and digital raster graphics.