Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi)

Paul C. Murphey, Robert P. Guralnick, Robert Glaubitz, David Neufeld and J. Allen Ryan*

*University of Colorado Museum of Natural History and *Department of Ecology and Evolutionary Biology, Bruce Curtis Bldg., 265 UCB, Boulder, Colorado 80309-0265, USA*

Received: 16 April 2004 - Accepted: 6 September 2004

Abstract

The vast majority of locality descriptions associated with biological specimens housed in natural history museums lack the geographic coordinates required for computer-based geographic analyses. Assigning such coordinates to existing specimen records is a process called retrospective georeferencing. The georeferencing of biological collections makes those collections more valuable by allowing them to be used in spatially explicit biodiversity analyses. Here we review some of the most common problems inherent to the retrospective georeferencing of biological collections. We then attempt to classify the most common types of locality descriptions according to our own rule-application for georeferencing, which was developed as part of a larger funded effort to create an online mapping and biodiversity analysis portal for the North-central Rocky Mountains and adjacent plains. As a means of comparison with our manual computer-assisted georeferencing method, we evaluate four currently available automated georeferencing tools. We argue that these automated tools are less accurate and precise, and less time efficient, than the three computer-assisted manual protocols for natural history museums: MaNIS, MaPSTeDI, and INRAM. Nevertheless, it is clear that these automated tools represent a critical first-step in the development of accurate, precise and efficient tools which will ultimately replace or at least enhance existing manual protocols. In order to facilitate the development of better georeferencing technologies, it is critical that workers involved establish a consensus set of locality types and rules for georeferencing them which will pave the way for the next generation of semi-or fully automated tools. A collaboratively built next generation georeferencing tool should be Internet-based.

Introduction

At least 2.5 billion natural history specimens exist in museums worldwide (Cotterill, 1995), with an estimated 500 million of these housed in US museums (Krishtalka and Humphrey, 2000). As collections grow and age, the challenges and costs of maintaining, conserving and repurposing for new research endeavors both specimens and associated documentation grow with them. These “libraries of life” represent an invaluable scientific resource and record of biodiversity change (Alberch, 1993), and many museums are now working in cooperation to leverage advances in information technology to maximally use this

biodiversity information. A major goal is to provide network access to vast amounts of collections-based data and the tools to make sense of that data (Brisby, 2000).

The term "Biodiversity Informatics" encapsulates this new synthetic discipline that integrates biological research, computational science, and software engineering to investigate complex evolutionary and ecological phenomena (Baker et al., 1998; Fischman, 1996; Grace, 1997; Krishtalka and Humphrey, 2000; Parker et al., 1998). Biodiversity Informatics applications have furthered the production of data standards and network protocols for sharing data (Stein and Wieczorek, 2004), and perhaps most importantly, provided new tools for visualizing and analyzing specimen records from multiple collections (Neufeld and Guralnick, in review). The end goal of all of these efforts is to meet the challenge of the biodiversity crisis by providing a scientific basis for the decisions facing society regarding the stewardship of life on earth (Wilson, 1992).

Despite the growing number of technological tools available to museum scientists and bioinformaticians, a vast amount of work remains to be done at the museum-collections level in order to prepare specimen data for analysis. It has been estimated, for example, that only 5% of natural history collections data are captured in electronic databases (Krishtalka and Humphrey, 2000; Beaman and Conn, 2003). Once databased, further work may be necessary to make the specimen record usable for further analysis. One crucial piece of information absent in most specimen records is a spatial coordinate from the locality information associated with the record.

The natural history museum community currently faces a major challenge of assigning spatial coordinates to locality data for hundreds of millions of older specimen records (Baker et al., 1998). Although advances in technology have made handheld GPS receivers a practical solution for the real-time capture of spatial coordinates when collecting in the field, text-based locality descriptions still remain essential and sometimes the only piece of information available. The process of converting a text based description into a geospatial coordinate is referred to as georeferencing. This georeferencing process is one that has just begun in earnest in museums, and the quality of the process directly determines the usability of the data in geospatial analyses.

The georeferencing process is ultimately a set of methods for overcoming the problems presented by the often poor state of locality data associated with natural history museum specimens. For tracking biodiversity change over time, our baseline information is from older collections which almost always do not have geospatial coordinates. Unfortunately, the older the collections data, the less accurate they generally are. This is partially because of changing geographic terminology and standards of collecting over time. It is also partially a legacy of inaccurate, imprecise, and/or vague locality coordinates or text descriptions recorded by the field researcher at the time of collection. Another significant factor is the disassociation of detailed records of collecting events (text, maps, photographs, coordinates) from specimens labels and the collections catalogue.

Many museums are now in the process of georeferencing their collections data with the common goal of making them available for geospatial analyses. This relatively new activity is leading to a combination of shared and independent development of methodologies and tools to expedite the process. If the goal is to have comparable georeferenced data across institutions, the development of different standards works against it. Therefore, best practices documents, like this paper, are meant to help with the creation of a single methodology for doing this important task.

Georeferencing is not just about assigning a point location, but also determining the confidence of the assigned point. If properly implemented, georeferencing increases the precision of raw locality data while transforming them into a comparable and thus analytically useful format. Because the degree of accuracy and precision rendered

by the georeferencing process is dependent on the raw locality data, a by-product of georeferencing should be the assignment of some measure of degree of error or confidence in the data. For the most part, those methods that currently include such confidence measures express and derive them differently.

Perhaps the most important aspect of georeferencing is the fact that it is a reproducible process. This fact is largely responsible for the gathering momentum behind georeferencing efforts in museums, because it means that the georeferencing process can transform locality data consistently and precisely. The importance of these efforts is underscored by their underlying purpose: to greatly increase the available body of data for biological distribution and diversity studies. The end goal is a quantitative and qualitative biological history derived from geospatial and temporal information unleashed from existing museum collections.

Thanks to a number of readily available and inexpensive computer-based geographic tools, the challenges presented by the georeferencing process are now finally less technologic than economic. The process of georeferencing, whether a computer-assisted record by record process ("manual georeferencing"), or using an automated batch processing tool, is time consuming. As we discuss in this paper, choosing the right georeferencing method is a compromise between precision, accuracy and speed. While one might presume that hand georeferencing is less efficient, batch processing also requires a substantial effort in data preparation and in post-processing quality checking. In deciding which current georeferencing method to implement, it is important to consider the amount of data to be coded and degree of precision desired versus the available resources. The more time that is devoted to the georeferencing process and the more rigorous the application of a clear rule-set, the more precise the results will be. Some automated batch processing tools are capable of rapidly georeferencing localities with relatively simple locality descriptions, but, as discussed below, the coordinates generated are often less precise and less accurate than those obtained by more time consuming manual computer-assisted georeferencing methods.

In this paper, we begin by discussing issues of accuracy versus precision with respect to field collecting and the retrospective georeferencing of museum localities. We then discuss problems and pitfalls in the georeferencing process, and propose a classification for the most common types of locality descriptions which is reflected in our rule-application for georeferencing developed at the University of Colorado Museum (UCM) for work on an NSF-funded distributed mapping project - Mountain and Plains Spatio-Temporal Database-Informatics Initiative (MaPSTeDI). We then evaluate four of the available automated georeferencing tools. We argue that manual, computer-assisted georeferencing methods are probably more reliable than current automated technologies, and not necessarily more time intensive. Finally, we discuss future development most likely to effectively and economically make biological collections data available for geospatial analyses through georeferencing. Detailed examples and treatments for georeferencing vague locality descriptions and applying locality confidence intervals are provided in appendices 1 and 2.

Accuracy and precision in the georeferencing process

Understanding the distinction between accuracy and precision is important both for the documentation of localities and the retrospective georeferencing of locality data. Both accuracy and precision begin to come into play at the time any locality is first documented. At that time, it is the responsibility of the field worker to accurately document the location where a specimen was collected, whether calculating distance and direction from a known point, plotting a position on a map, or recording coordinates generated by a GPS receiver. Accuracy at the point of collection thus

refers to the quality of the locality data originally reported by the field worker.
At the time of collection, a field location is either accurately or inaccurately described, no matter what level of geographic detail is recorded. A locality can be inaccurately described by the original field worker, but then later accurately and precisely georeferenced. Once a locality has been documented by a field worker, its original accuracy or inaccuracy is maintained if additional error is not introduced during data entry or other processing, including georeferencing. If a mistake is made, an accurately documented locality can be rendered inaccurate, and this may go undetected. Accuracy in georeferencing involves correctly positioning locality points based on available locality data, and correctly entering the data into the georeferencing database.

Precision in the context of georeferencing refers to the potential amount of geographic extent represented by the locality. Precision is also an outcome of the original locality description and application of georeferencing method. Note that a locality description that lists only the county in which it is located may well be accurate, but because a county is a large geographic area, it is relatively imprecise. On the other hand, a locality which lists x, y UTM coordinates with one number transposed is relatively precise because only a small geographic area is involved, but inaccurate because the UTM was erroneously transposed. In this case, it may be impossible to tell which number was reported incorrectly, and to determine where the original locality was.

The georeferencing process often increases locality precision. During georeferencing, locality descriptions are examined, evaluated and geospatial coordinates are assigned to them. When this process results in a reduction of the size of the geographic area encompassing all possible locations for the point (a decrease in the size of the potential area of geographic error), it increases the precision associated with the locality. For example, the locality "2.5 miles west of Boulder Falls" was determined to have been 2.5 miles along Boulder Creek, and not due west of the falls "as the crow flies." This finding increases the precision of the locality record. Arguably, generating x, y coordinates for text-based locality descriptions always increases precision because the process transforms text into a format which is suitable for analysis, makes it directly comparable with other georeferenced localities, and generates an estimate of error.
In summary, the goal of the georeferencing process is threefold: 1.) To assign computer readable coordinates for location and error and by doing so; 2.) To maintain or increase the accuracy of the original locality record by not introducing error and by correcting erroneous locality data; 3.) To maintain or increase its precision by evaluating and refining the available locality data in order to assign geospatial coordinates.

Methods and Discussion

General difficulties with the georeferencing process

Locality descriptions in most collections databases and paper catalogues vary greatly in the quantity and quality of geographic information they contain, and pose a number of problems which manual and semi/fully-automated georeferencing methods are currently seeking to overcome. These problems are the result of: 1) Text-only locality descriptions without geographic coordinates; 2) Inconsistent formatting and misspellings; 3) Older records containing place names that have since changed location and/or are no longer in use; 4) Vague, imprecise, confusing and/or contradictory locality descriptions, and various combinations thereof. Furthermore, a significant challenge is posed by the fact that much of the highly accurate and precise locality information associated with collections objects is not directly associated the specimen or locality records, catalogue numbers, or paper labels. Much of this disassociated ancillary locality information can often be found in collections of field notes, maps, photographs, and other archival materials.

A major problem with georeferencing is that accuracy and precision of geographic data recorded with the specimen record generally decreases with the age of the record. The most obvious example of this is the lack of geospatial coordinates associated with most records about 40 years or older. For example, only 4% of the paleontological localities at the UCM recorded prior to 1960 include geographic coordinates (e.g. Public Land Survey System [PLSS], State Plane, Latitude/Longitude, or Universal Transverse Mercator [UTM]). For the year 2001, 98% of the localities recorded in the same database included UTM coordinates, and were thus essentially already georeferenced.

In addition to a lack of coordinates, the vast majority of older locality records in collections databases consist of unformatted text descriptions along with occasional spelling errors, which typically reference the locality to some identifiable geographic feature such as a town, mountain or river. The "georeferencer," whether a human being or computerized tool, must be able to recognize place names, make sense of the confusing array of formatting inconsistencies including word order, abbreviations and punctuation, and be able to distinguish and prioritize parts of the locality description that should be given more weight in the assignment of coordinates. Such prioritizations are still difficult for a computer to perform, as they may be as obtuse as knowing, for example, that a particular field worker preferred to collect on one side of a valley rather than the other. This information, which could be obtained from archival field notes, is much more likely to be discovered by employing "manual" computer-assisted georeferencing methods rather than using one of the existing fully automated tools.

Vague, imprecise, and contradictory locality descriptions are major problems when working with older records. These problems stem mostly from recording errors either at the time of locality documentation or data entry, place name changes through time, and the changing standards of acceptable precision made possible by better maps and geographic technologies that were unavailable to most field workers until the last several decades. As an extreme example, what may have been considered an adequate and precise locality description in the nineteenth century, such as "North Platte River, Wyoming Territory," was perfectly acceptable for the research standards of the day, but would usually be considered inadequate by modern standards. Obviously, locality descriptions with more detail, even if no x, y coordinates are included, have greater georeferencing precision potential than more vague and/or imprecise descriptions, although more detail can also lead to greater chances for inconsistencies. The least precise locality descriptions are referenced to large geographic features such as cities, counties, states, and even countries or continents. These features have such large geographic extents, that they are generally not useful for positioning a locality with adequate precision for meaningful analysis.

A surprisingly high number of older locality records contain references to place names for ranches, towns, roadways, watersheds, mountains and other anthropogenic and non-anthropogenic features which have since been renamed or eliminated. This can result in contradictory and/or confusing locality descriptions. Georeferencing such records requires access to historical place names data in gazetteers, government records, and/or field notes. Place name discrepancies within locality descriptions, such as counties which don't contain a specified town, for example, should be an indication to the georeferencer that a change or elimination of a place name has occurred.

Other problems presented by the passage of time are the result of changes in the physical locations or in the extent (size) of features. The latter is of particular concern with cities and towns which have grown in size since they were used as the geographic reference point for a locality. For example, when a locality description is given as "10" miles north of a certain town, it is usually assumed

that the approximate center of the town (usually at the post office) was used as the reference point. In the intervening years, however, the size of the town may have increased dramatically, and its center may now be in a completely different location. Georeferencing procedures should take into consideration the date when the locality was first recorded, so that the contemporaneous extent and location of the reference point can be more accurately and precisely estimated. One relatively common example is change of position of county boundaries subsequent to the original recording of a locality. This causes confusion because the locality description may include features that appear to be in the wrong current county. Historical documents showing demographic and political boundary changes are essential elements for correctly geocoding older records. It should be noted that place name databases like the Geographic Names Information System (GNIS – discussed in detail below) does not necessarily position place name coordinates in the exact center of a feature, especially in the case of linear features such as rivers, in which the point of origin is usually used. GNIS also contains rare errors.

It is not uncommon to encounter duplicate or similar place names in locality descriptions of any age. Dry Creek, Little Dry Creek, and Big Dry Creek, for example, could all occur within several miles of one another, and some earlier workers may have referred to all of them simply as Dry Creek. Multiple "Bear Lakes" occur within the same mountain range. Other examples abound. The possibility that duplicate or similar names for the feature named in the locality description occur in the same general geographic area should be investigated, especially when the place name is somewhat generic.

The MaPSTeDI georeferencing protocol

A major goal of the MaPSTeDI method is to retrospectively georeference collections data with the highest possible degree of accuracy, precision and efficiency. In order to meet the challenges discussed above with the variety of locality descriptions, textual and geospatial, we established rules for georeferencing different types of localities based on a classification of localities (Table 1), and then implemented a method for georeferencing based on these rules. The end result is a computer-assisted manual georeferencing protocol that produces consistent and reproducible results, and that works for locality data associated with zoological, botanical and paleontological specimens housed in participating MaPSTeDI institutions.

The development of an accurate and precise georeferencing method was the first major objective for MaPSTeDI because it generates the necessary computer-readable geospatial coordinates that forms the basis for the examination of distribution and biodiversity changes for the later parts of the project – the online GIS application (http://www.geomuse.org). The MaPSTeDI georeferencing method is simple to learn, and when implemented correctly, is capable of processing the most complex locality descriptions. It includes the following steps which are discussed in detail using locality examples below:

1) Finding locality points and assigning coordinates
2) Assigning locality confidence values
3) Recording data and documenting georeferencing rationale
4) Flagging records for further review if necessary
5) Quality checking

Preparation and Setup. Implementation of the MaPSteDI method required customizable database software, topographic map software, Internet access, access to a gazetteer, and all available collections archives. Using three to four part-time undergraduate student "georeferencers," a part-time graduate student quality checker, as well as input and oversight by other museum personnel, approximately 60,000 UCM records were georeferenced from September 2001, to September 2003, at an average of 5 minutes or less per record. This rate includes both

Table 1. A classification of locality types from generally most (top) to least precise (bottom).

Most Precise Locality Data Type		Examples of Locality Descriptions
Latitude/Longitude or UTM Coordinates		Boulder Falls, 465408 mE, 4428396 mN, Boulder County, Colorado
		Upper Red Rocks Canyon, Piñon Canyon Maneuver Site, UTM 603960/4156010. T29S R57W, NE_ SE_ NE_ sec 6., Las Animas County, Colorado
		Lat. 38º43'23", Long. 102º39'39", Cheyenne County, Colorado
Public Land Survey System, or Cadastral (Section, Township, Range)		Boulder Falls, PM 6 T1N R72W Sec.35 NW_ NE_ NW_, Boulder County, Colorado
		2.5 miles S Las Animas, in the Purgatoire State Wildlife Area. SW_ NW_ Sec 27, T. 23 S. R.52 W.
Place/Feature Names		
	Waterfall	Boulder Falls, Boulder County, Colorado
	River	Arkansas River, near Pueblo, Colorado
	Lake	Baseline Lake, Boulder, Colorado
	Address	RR 7, Box 539A, Clear Creek County, Colorado
	Building	Jefferson County Courthouse, Colorado
	Roads/Highways	SH 50/Interstate 25, El Paso County, Colorado
Offsets*		
Single Offset	Determined by Linear Feature (creek)	2.1 miles NW of Boulder Falls, Boulder County, Colorado
Double Offset	Determined by Linear Feature (road)	20 mi NW Delta, 3 mi up Dominguez Canyon, Mesa County, Colorado
Single Offset	Determined by Bearing	6 miles SE Boulder, Marshall Reservoir, Boulder County, Colorado
Double Offset	Determined by Bearing	10 mi W and 7 mi S of Ninaview, Las Animas County, Colorado
Triple Offset	Determined by Bearing	6.5 Km east, 1.5 Km north and 1.2 Km northeast of Agate, Elbert County, Colorado
Feature Name Issues		
	Duplicate or Similar Names	Dry Creek, Wyoming
	Name Changes and Updates	Highway 40, 2 miles west of Idaho Springs, Colorado = Interstate 70, 2 miles west of Idaho Springs, Colorado
	Name Eliminations	Sanchez Ranch, Costilla County, Colorado
Vague/Imprecise Locality Descriptions		
	Large Area	Pitkin County, Colorado
	Large Feature	Sangre de Cristo Mountains, Colorado
	Elevation	Altitude 6400', Boulder County, Colorado
	Linear Feature	Colorado River, Eagle County, Colorado
	Vague Distance	Near Boulder Falls, Colorado
	Vague Direction	5 miles from Boulder Falls, Colorado
Contradictory Locality Descriptions		
	Contradictory UTM Zone and Coordinates	11S, 292638 mE, 4382112 mN
	Contradictory Meridian and STR	PM7 T1N R72W Sec.35
	Contradictory Place/Feature Name and County	Dillon Lake, Las Animas County, Colorado

*Multiple offsets do not necessarily result in greater precision.

quality-checking and dealing with problem records as discussed below. In sorted data sets with few problem records, the rate is significantly faster. As with automated methods, significant time savings can be achieved by preparing data for processing prior to georeferencing (e.g. formatting, spell checking). We found higher efficiency when the data were sorted so that records with similar (or identical) locality information are grouped. For example, sorting by state, county, township, and even section if possible significantly increases the pace of georeferencing because it minimizes the amount of time spent navigating computerized topographic maps in order to plot locality points.

The georeferencing aids that were employed included computer and paper maps, place name databases and various newly created and existing databases for storing collections and georeferencing information. Important features for topographic map software included seamless 7.5' USGS Topographic Quadrangle base maps, user friendly panning, navigation, measuring and distance tools, multiple assignable waypoints, and multiple and easy-to-switch map datums, coordinate formats, and distance units. The ability to search for section, township and range is very useful, but is not a common feature in products we tested. We used both National Geographic Topo! and MapTech Terrain Navigator. The latter is capable of searching for section, township and range, but we preferred the display and overall features of Topo!. The North American Datum of 1927 (NAD27) was preferred because of its relatively high accuracy when used with USGS 7.5' Topographic Quadrangle maps. Paper atlases were often used to supplement the digital maps.

The primary digital gazetteer used by MaPSTeDI is the USGS Geographic Names Information System (GNIS). This gazetteer is widely used and available online (http://geonames.usgs.gov), or on CD, and can be easily incorporated into the georeferencing database for more efficient data retrieval. The use of a widely available resource like GNIS is particularly desirable because it permits direct comparison of georeferenced data from different databases and institutions. For additional information about the MaPSTeDI GNIS usage, see http://mapstedi.colorado.edu/documents/fm_gnis.html. Other place names and geographic information sources are numerous, and included specialty gazetteers, websites, historical gazetteers, electronic databases, and archival field notes and maps.

The MaPSTeDI method used the following data fields which were created for each specimen record in a work copy of the collections database.

UTM Zone
UTM Easting
UTM Northing
Township (when available)
Range (when available)
Section (when available)
Confidence Value
Georeferencer (name of person initially georeferencing the record)
Date of Georeferencing (date of initial georeferencing)
Quality Checker (name of last person to quality check the record)
Date of Last Quality Checking (date of last quality check)
Place name data source (gazetteer name, only used if a gazetteer is used)
Georeferenced (yes/no, used for easy extraction of georeferenced records)

Record progress (includes a text description of the georeferencing rationale for each locality)

The MaPSTeDI method georeferenced each specimen record separately, instead of creating a separate table of distinct localities. Our reason was that temporal information and species data should both be taken into account when initially georeferencing records. Many records have similar localities that could change meaning depending on the collection date. For instance, a locality reading "just north of Colorado Springs town limit" would be placed in one location in 1900 and an entirely different location in 2000. Species

information is also useful for initial georeferencing, not just validation. It is unlikely that a cactus would be collected in a marsh or a fish would be collected on top of a rocky hill. However, by georeferencing just localities, errors like these can occur.

The georeferencing information was entered into a copy of the original collections database. This precaution was particularly important since georeferencing was not being done exclusively "in house" by collections staff. Instead, georeferencers were hired who worked with both the MaPSTeDI team and collections staff. When modifying the georeferencing databases, we almost never overwrote original data but instead would only add data to new fields. Obvious spelling errors were corrected, but the correction was also noted in the record progress field. The integrity of the original data was always maintained in order to ensure that additional error was not accidentally introduced during georeferencing, and to preserve both the raw and georeferenced versions as documentation of the process.

The MaPSTeDI georeferencing protocol was designed for collections of mostly recent terrestrial and freshwater organisms. Locality data associated with such collections typically include some sort of x, y coordinate or two dimensional information, but older records commonly lack direct measurement of elevation or depth data (z = vertical axis). Although it is outside of the scope of this paper to discuss the various ways z-axis information can be assigned to locality records, it should be noted that for other types of collections like marine and paleontological material, the "z" axis is often crucial to assign when the specimen is data is collected, while that has not been as much the case in recent terrestrial collecting. However, it is also increasingly common and valuable for biologists focusing on terrestrial and freshwater organisms to access depth/elevation data for specimen records.

Finding Locality Points and Assigning Coordinates. The first step to georeferencing is the process of "finding" each locality on a map using the existing locality description in the database, and assigning geographic coordinates to it. We used the topographic map software packages described above to locate the point based on the locality description and then generate x, y coordinates in Latitude/Longitude or UTM format in user-defined datums. Even if localities were already plotted on archival paper maps, topographic software made it easy to calculate distance and direction between the locality point and other features referenced in the locality description, and was much faster and more accurate in the assignment of coordinates than working with a hard copy map.

MaPSTeDI employed the UTM (Universal Transverse Mercator) coordinate system to record geospatial coordinates. While the use of Latitude/Longitude remains the current standard amongst most georeferencing projects, UTM coordinates were another logical choice. All major topographic software packages display coordinates in both UTM and Latitude/Longitude formats, although National Geographic's Topo! displays UTMs more precisely than Latitude/Longitude. We found UTMs easier to record because the georeferencer does not have to keep track of plus and minus signs, as well as decimal places. Additionally, UTMs are simple for georeferencers to work with because each unit represents one linear meter. Offsets can therefore be easily calculated by the linear distance between two points. One problem with the use of the UTM coordinate system is the reliance on UTM zones, which can render the coordinates useless if they are absent, inaccurate and/or lack additional locality information.

We established a classification of locality descriptions shown in Table 1. The classification describes the different kinds of locality descriptions that are present in the databases we georeferenced for MaPSTeDI. Georeferencers matched each record against the classification scheme and applied a different rule-set for

georeferencing depending on the classification of that record. The different kinds of classifications and rules for georeferencing are discussed more fully below, although not all permutations shown in Table 1 are covered.

Latitude/Longitude or UTM Coordinates. Even if x, y geospatial coordinates are already recorded for the locality either by hand or by GPS unit, we found it important to verify that these coordinates match the rest of the locality data if there are any. If the locality description and coordinates match (Figure 1A), it is likely that the coordinates were recorded in the field with a GPS unit, correctly assigned using a paper map, or looked up in a gazetteer. Discrepancies can occur if either the coordinates or other parts of the locality data are inaccurate (Figure 1B). In the example shown in Figure 1B, it appears as if the Latitude/Longitude coordinates were incorrectly recorded because they don't precisely match the location of Boulder Falls. Other sources of error include rounding coordinates when using a UTM grid on paper maps in the field (Figure 1C), failure to record the correct datum associated with coordinates, and GPS error, which can be as great as 200 meters in older units. With selective availability (SA) turned on, the x, y coordinates will fall within 160 m of the actual coordinates about 95% of the time (see http://www.ngs.noaa.gov/FGCS/info/sans_SA/).

If the coordinates provided do not match the rest of the locality description, we carefully evaluated the discrepancy in order to determine whether the locality description was more accurate than the coordinates given, or vice versa. If the locality was described as "Boulder Falls" but the coordinates accompanying that description were not the precise location of Boulder Falls (Figure 1B), a degree of error was assigned that encompasses both, regardless of which one is selected as the georeferenced point.

Township, Range and Section (Public Land Survey System). One major problem with PLSS coordinates is the order of sequence of quarters within a section, and sections within Townships, which, although standardized, is not correctly applied by all workers. Furthermore, important differences exist between the PLSS in the United States and the similar Dominion Land Survey system used in Canada. If PLSS (cadastral) coordinates are the most specific and precise (or the only) information provided in a locality description, the locality point was placed in the center of the section, quarter section, or smallest division thereof. If other locational data were provided, such as a place name, that information was compared to the PLSS coordinates in order to determine whether they match (Figure 2A). If they don't (Figure 2B), the important question became what portion of the locality description is most accurate? Did the field worker use a topographic map in the field, or were the PLSS coordinates added to the record later? It is possible that they were added even decades later, and potentially imprecisely or inaccurately by another person working with the collections data. Note that it would be far less likely that the PLSS coordinates were assigned by the original worker, and place name information added later. Furthermore, a more precise locality description combined with precise (_, _, _, _ Section) PLSS coordinates is a good indication that the field worker was conscientious in the collection of locality data, but does not necessarily indicate that the worker described and/or plotted the locality accurately (i.e. knew where they collected the specimen(s) and reported that location accurately).

Place Names. The majority of locality descriptions will reference a place name with or without additional modifiers like distance offsets (Figure 3A). Assigning coordinates for such localities is usually simple, especially if the place name is present in a database such as the GNIS, which includes almost two million place names in the United States. If the place name is not found in GNIS initially, it is often useful to check alternate spellings and partial names. If the place name still cannot

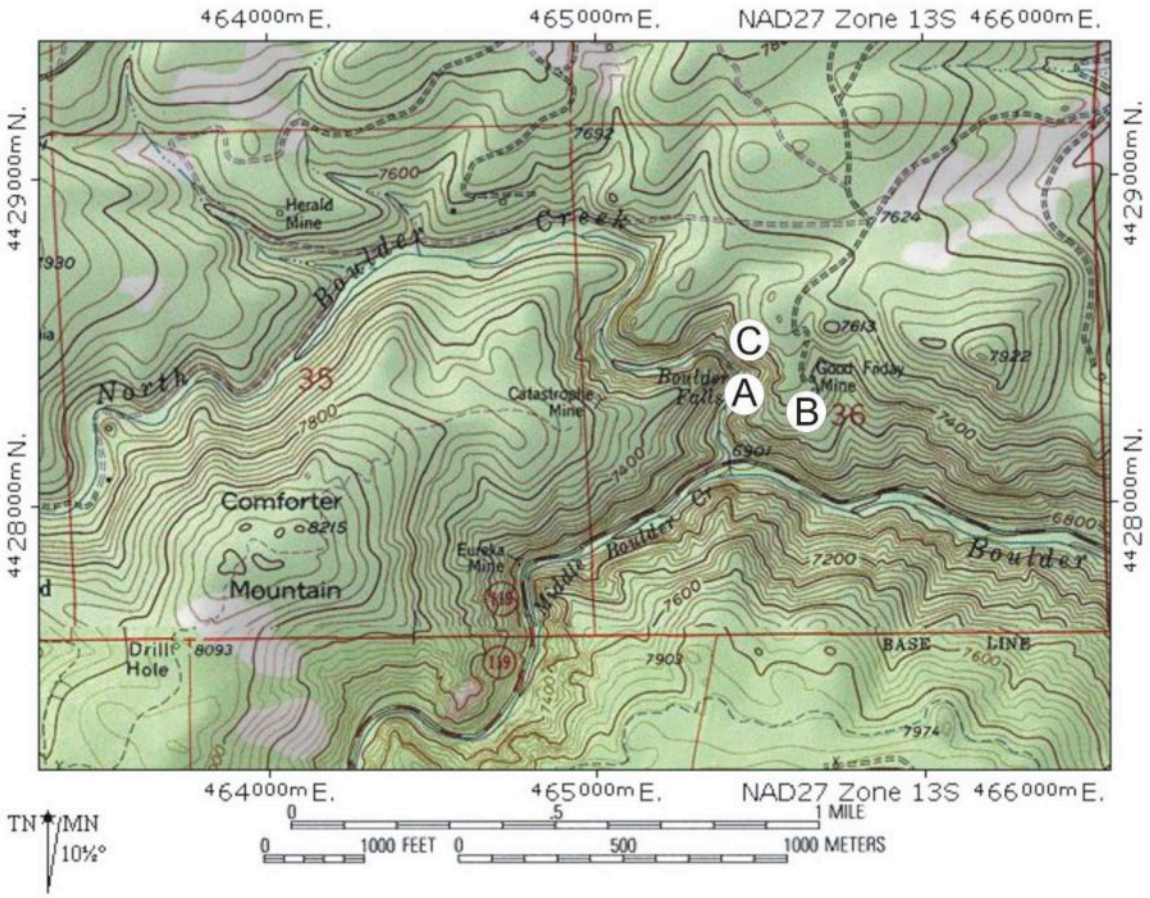

Figure 1. A, Boulder Falls, 465408 mE, 4428396 mN, Boulder County, Colorado; B, Boulder Falls, 40º 00’ 23’’ N, 105º 24’ 11’’ W, Boulder County, Colorado; C, Boulder Falls, 465500 mE, 4428500 mN, Boulder County, Colorado.

be found, the georeferencer searched additional resources including the Internet and historical gazetteers. Georeferencing localities consisting of place names is more difficult if the locality description contains conflicting information, such as an incorrect county, for example. If additional information is included with the place name, the position of the locality point should be adjusted accordingly (Figure 3B).

Offsets from Place Names. Many locality descriptions include a reference to a geographic feature using some measure of distance and direction. This reference is referred to as an offset. In the United States, offsets typically include a compass bearing and distance in statute miles. An example of a single offset locality is “6 miles East of Boulder, Colorado.” Double offsets (e.g. “6 miles East and 3 miles South of Boulder”) and even triple offsets also occur (Table 1). Locality offsets often represent estimates made by field workers, so it should not be assumed that a map or compass was used. “Linear feature” offsets are those which were referenced to a place by measuring the distance along a feature such as a road or river, which are usually not perfectly straight, but are linear. “Bearing” offsets are those which were referenced to a place using one or more directions, either estimated or determined using a compass or other navigational device.

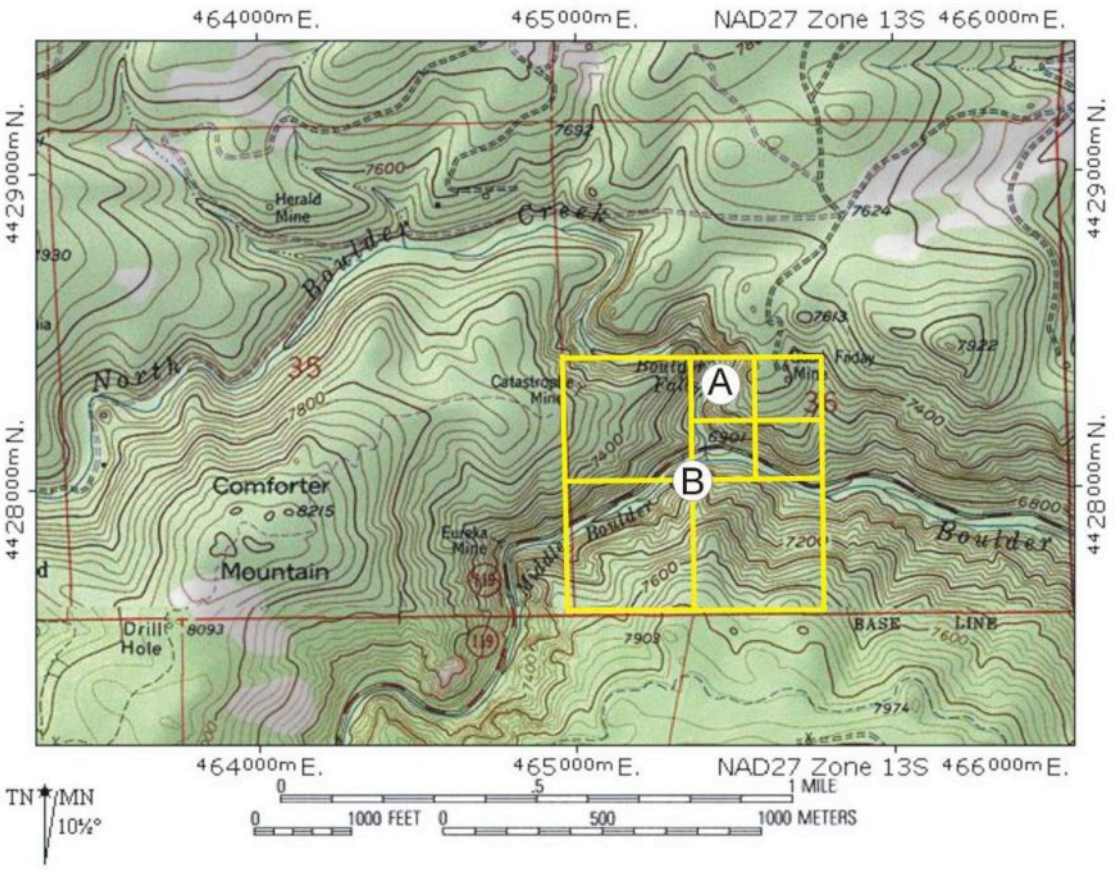

Figure 2. A, Boulder Falls, PM 6 T1N R72W Sec. 36 NW_ NE_ SW_, Boulder County, Colorado; B, Boulder Falls, PM 6 T1N R72W Sec. 36 SW_, Boulder County, Colorado.

Depending upon the content of the locality description, it may be possible to distinguish between offsets determined using a "linear feature" from those determined using an estimated directional bearing ("as the crow flies"). In Figure 4A, the offset was most likely determined by air (directional bearing) since there are no linear features that run northwest from Boulder Falls. The GNIS coordinates for Boulder Falls are used as the origin of a line that measures 2.1 miles directly northwest from the falls, and the georeferencer plotted the locality at the end-point of that line. Note that multiple directions (offsets) in locality descriptions almost always represent bearing offsets (Figure 4B).

We considered linear offsets to be more likely than bearing offsets when trails, roads, rivers or even ridgelines were likely followed to get from Point A to B. For example, the distance one travels by hiking 2.5 miles west from Boulder Falls along Boulder Creek is different than 2.5 miles due west, because Boulder Creek is not straight (Figure 4C), meandering slightly North, then West, and then Southwest. The fact that the upstream Boulder Creek drainage is generally west of Boulder Falls raises the possibility that the field worker hiked along the creek to determine the length and distance of the offset. This assumption could be strengthened if the specimen collected was, for example, a fish, and no other streams were mapped in the immediate area. In this case, the GNIS coordinates for Boulder Falls are used as the origin of a line which is traced 2.5 miles west along North Boulder Creek. Linear offsets may result in more precisely positioned localities, because it is only necessary to measure along the linear feature for the reported distance from the reference point to plot the locality (i.e. the x or y coordinate is already known). In the

absence of more precise information about starting points, we measured from the GNIS starting point for the referenced place name.

In summary, if the georeferencer can establish whether the worker's locality offset was determined by linear feature or by using a directional bearing, the precision of the coordinates assigned will be increased by the georeferencing process. In some cases, it is not possible to evaluate whether an offset was measured with a bearing or by linear feature. If it is not possible to make a determination of what type of offset was used, the georeferencer should plot the locality point midway between the two possibilities and assign the appropriate confidence value. Note also that cases in which 1/10 of a mile is used in the locality description (e.g. 2.1 miles) suggest that care was taken to accurately and precisely record the location of collection, especially in comparison to, for example, "about 2 miles west."

Offsets associated with streams and rivers often include the descriptors "above" and "below," instead of, or in addition to, cardinal directions. Above is used for the upstream direction, and below for downstream. The direction a river flows can be easily determined on a topographic map by looking at the contour lines and elevations since contour lines will always point upstream as they cross the river. We also used caution when georeferencing localities associated

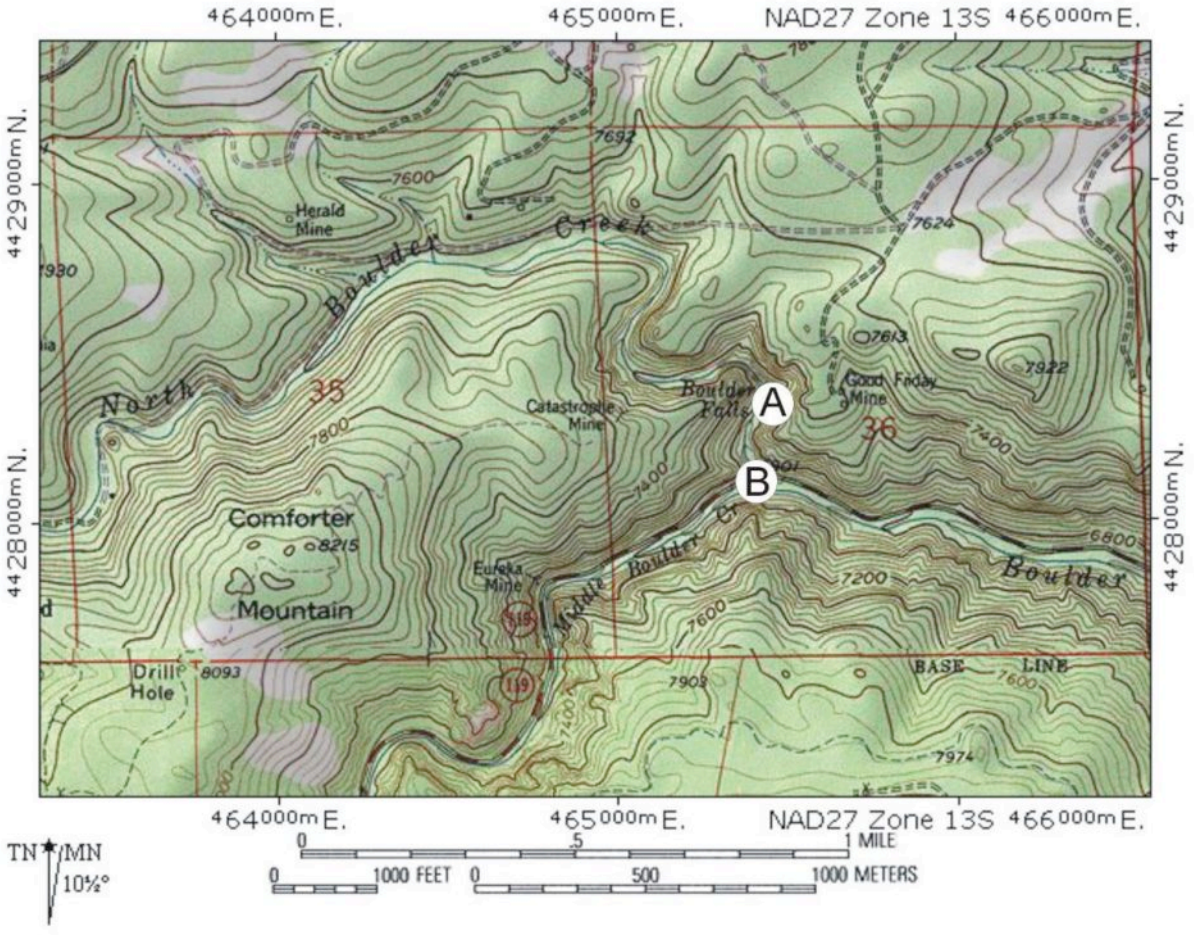

Figure 3. A, Boulder Falls, Boulder County, Colorado; B, Just below Boulder Falls at confluence with Middle Boulder Creek, Boulder County, Colorado.

with aquatic taxa because streams and lakes change size and location, and may have moved since the locality was first recorded.

Other Locational Modifiers. Other types of information were often included in locality descriptions, and we considered these during the georeferencing process. Some include elevation, date of collection, highway and street identifiers, addresses, building names, and linear features with no accompanying information. Elevations included in locality descriptions were always treated with caution because they are frequently estimated either by the collector or later by data stewards, especially in older records. For this reason, when used in combination with other locality data, they often create inconsistencies in the positioning of the locality point. Among the most reliable of elevation data are those obtained by the field worker who fixed her position on a relatively small scale topographic map with 20' contour intervals in topographically complex terrain. Most currently available handheld GPS receivers yield inaccurate and inconsistent elevations. If the MaPSTeDI georeferencers knew that elevations were determined using a small scale topographic map or GPS, the elevation information can be useful for records which include only elevation and a linear feature such as a road or stream (e.g. North Boulder Creek at 7,500 feet).

Road maps and Internet resources such as , http://www.topozone.com/, and http://www.mapquest.com/, were useful in georeferencing localities consisting of street addresses, but addresses also change over time. Building names are also sometimes used in locality descriptions. Very few are included in the GNIS database, but many can be found in the local yellow pages. Buildings, streets and addresses are subject to more frequent name changes than natural features such as rivers, lakes and mountains, and in some cases it may be necessary to find a record of their location at the time the locality was documented in order to accurately georeference the record. References to historical buildings and roads may be found on the Internet, often in specialized historical websites focusing on changes to towns and roads. County clerk offices often have good historical records of old business names and locations and can be consulted.

Highway numbering and naming systems were often used in locality descriptions and led to confusion among georeferencers. The highway numbering system for interstate and U.S. highways consists of odd numbering for north-south highways and even numbering for east-west highways. In reality, however, most are not straight, and many do not travel perfectly in the direction indicated by their numbering. Furthermore, the system of even and odd numbering applies only to highways with one or two digit route numbers, and often does not apply to state highway systems. The Santa Cruz Public Libraries website provides a short description of interstate and US highway numbering rules at http://www.santacruzpl.org/readyref/files/g-l/hiwaynos.shtml. Like road names, highway numbers also change over time. All this information was considered when locality descriptions contained such locational modifiers.

Obviously, vague locality descriptions are the most difficult to georeference, because they lack the information to needed to plot the locality with an acceptable level or error or confidence. Vague information is not necessarily inaccurate, but is imprecise. Vagueness may be the result of data which was not correctly or fully entered into the database and the original hard copy of the collections catalogue should be checked against vague databased locality record to eliminate this possible source of error. Common examples of vague locality descriptions and suggestions for georeferencing them are provided in Appendix 1.

Assigning Locality Confidence Values. Because of the uncertainty inherent in retrospective georeferencing, it is necessary to generate a measure of error that

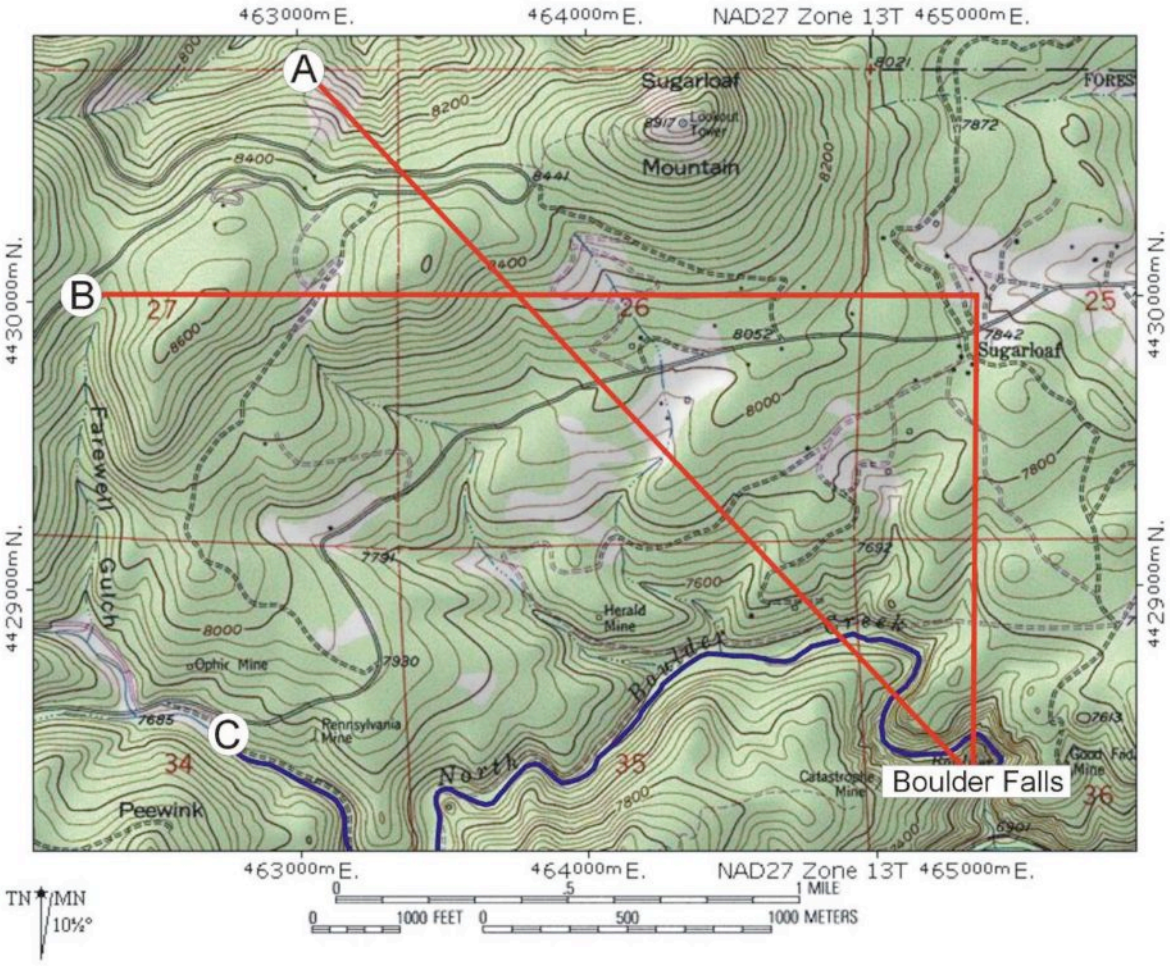

Figure 4. A, 2.1 miles NW of Boulder Falls, Boulder County, Colorado; B, 1 mile north, 2 miles west of Boulder Falls, Boulder County, Colorado; C, 2.5 miles W of Boulder Falls, Boulder County, Colorado.

accompanies the processed record. MaPSTeDI utilizes a system based on confidence values (CV's) on a scale of 1 to 9, with 1 representing the highest degree of confidence, and 9 the lowest (Table 3). This scale reflects the confidence in the geospatial precision of the locality coordinates based on the amount and type of data provided in the locality description. It does not reflect an estimate of the accuracy of the unprocessed or processed record. Thus, a CV of 4 means that the georeferenced locality point is considered to be within a 1 km radius of the location of the original locality based on the best available information (collections database, bound catalogue, field notes, etc.). Like all error indicators, confidence values are assigned after assigning coordinates to a locality, and must be based on the content of each locality description. Examples of confidence value assignments are provided in Appendix 2.

We developed our method of confidence estimation independently from other georeferencing projects which were concurrently developing methods. Two other approaches to estimating locality error in georeferencing include the point-radius and shapefile methods. The point-radius method (Wiekzorek et al., in press) represents error by measuring the radius of a circle that encloses the error region. The shapefile method (http://www.calacademy.org/research/informatics/georef/) creates a polygon coverage in

Table 3. MaPSTeDI Confidence Values.

1	Exact coordinates (UTMs or Lat/Long to seconds or decimal minutes provided, and no reason to suspect their inaccuracy)
2	Amended exact coordinates (UTMs or Lat/Long provided but were not accurate)
3	Public Land Survey System coordinates (_ Section, Section, Township, Range, Meridian)
4	Within 1 km radius
5	Within 5 km radius
6	Within 10 km radius
7	To County, or greater than 10 km but less than County
8	To State
9	To project region (or country)

ArcView to project an exact shape of the locality error visually on a digital basemap (a circle is used as the default error shape). Given the competing methods for error estimation, we consider it valuable to discuss the alternatives and highlight some of the reasons we decided on our approach.

An ideal method for estimating locality error would be able to, as precisely as possible, represent three general types of localities:

1. Localities with errors that are circular or approximately circular (points, bearing offsets, towns and cities).
2. Localities with errors which are of complex shape, typically not circular, and that are relatively large (mountain ranges, counties, countries).
3. Localities with linear errors (linear offsets, features such as roads, rivers, ridge lines, etc.).

The great advantage of using shapefiles to represent locality error is that they can be used to precisely represent error associated with any shape and size of locality. Disadvantages include the time required to digitize locality error for each locality, the difficulty of making comparisons between locality errors because of the complex shapes potentially involved, the necessity for georeferencing personnel to be trained to use GIS software, and the storage requirements for the digital basemaps and shapefiles. The point-radius method works well for most localities and permits quantitative comparison between locality error. It is not as precise for localities with linear error such as roads or rivers, or non-circular error such as most counties, and is more complex for the georeferencer to calculate than the assignment of the MaPSteDI confidence value.

MaPSteDI confidence values are relatively simple to calculate and permit easy comparisons between locality error values in the dataset, but they are not as precise as either the point-radius or shapefile methods. The MaPSteDI method was developed with 9 confidence values, but could be more precise if more were values were used to represent more shapes and sizes of locality error. As georeferencing efforts continue at institutions around the world, we expect that new approaches to the problem of locality error will be developed and existing approaches refined. As with many issues in georeferencing, the best solution is a tradeoff between precision and efficiency. Right now, the point-radius method represents a middle-ground solution between MaPSTeDI's category approach and shapefile solutions.

Recording Georeferenced Data and Documenting Georeferencing Rationale. Most of the mistakes made during the georeferencing process are the result of data-entry errors. In addition to accurately entering data, we made sure MaPSTeDI georeferencers were trained to fill in all required database fields as completely as

possible using the correct format. The database administrator placed constraints on some fields to force the desired format. Once georeferenced, datasets were processed with validation scripts or macros to ensure any formatting mistakes were corrected.

To minimize errors, promote consistency, and maintain data integrity throughout the georeferencing process, the following procedures were used in MaPSTeDI and can be recommended here: 1) Data generated by the georeferencing process should supplement the raw data, not replace or modify it; 2) Metadata generated by the georeferencing process should include the georeferencing decision making progress, or justification, for each record (this effort can be minimized if data are sorted for georeferencing); 3) Problem records should be flagged and then georeferenced by the most experienced personnel; 4) Quality checking should be performed on a random set of records in order to verify that georeferencing procedures are being followed; 5) All georeferencing personnel should be thoroughly trained; 6) Georeferencing personnel should communicate via a daily log to identify problems and solutions, and promote consistency.

Documentation of the georeferencing rationale for each locality was an important step in the MaPSTeDI georeferencing method because it creates a record of the decision making process which can be evaluated by future researchers. For problem records, including those which contain confusing, complex, or unusually detailed locality information, this documentation also permits quality checking personnel to understand the georeferencing decision and greatly expedites the quality checking process. We considered it vital that the georeferencing rationale be attached to each georeferenced record in its own database field to prevent loss of this important metadata.

The MaPSTeDI georeferencing method emphasized the importance of communication between georeferencers. Because georeferencers are commonly working on similar geographic areas within the six state project area, the daily logs of other georeferencers often provided clues that assisted with problem records. Conversely, it permitted more experienced georeferencers and quality checking personnel to spot potential georeferencing problems with less experienced personnel, and thus helped promote accuracy and consistency.

Flagging records for further review. For problem records, which typically consisted of complex or confusing localities, it was sometimes necessary for georeferencers to mark records for further review by more experienced quality checking personnel. For the MaPSTeDI project, problem records were flagged in the database and identified in the daily log. Many problem records were older, and commonly had locality descriptions which did not correspond to modern maps and were not listed in the GNIS. To georeference these localities, it was often necessary to consult a variety of information sources (e.g. original museum catalogues, scientific and other publications, archival notes, field maps, historical maps, specialty gazetteers, Internet websites, historical photo collections, genealogy databases, the yellow pages, and even collections personnel who were most familiar with the collections data).

Quality Checking. The development and implementation of a quality checking program helped identify problems and correct errors in the georeferenced data. The MaPSTeDI project used the following system: For newly trained georeferencing personnel, the first 200 records were checked as a supplement to the training process, and georeferencers learned by making mistakes. Following the initial 200, an additional 100 records were checked if necessary. Once trained, quality checking was reduced to 10 randomly selected records out of every 100 completed. If more than 2 of the 10 records were incorrect, an

additional 20 were checked. If additional errors were found, the entire 100 were checked. For experienced personnel, quality checking was reduced to 5 records for every set of 100.

Evaluation of georeferencing methods – automated and manual

An increasing number of institutions are currently developing and employing new georeferencing tools and methods. We processed biological collections data from the University of Colorado Museum in order to compare four of these. All are free, and are well documented in the literature and/or the Internet:
Biogeomancer (Beaman and Conn, 2003), University of Kansas
MANIS Georeferencing Calculator, University of California at Berkeley
GEOLocate, Tulane University
ArcView Georeferencing Extension, California Academy of Sciences

Our evaluations were conducted using three sets of herpetological collections data. The first (Set A) consisted of 83 randomly selected records, and the second (Set B) of 35 records which were selected because they contain more complex locality descriptions with multiple offsets, combinations of coordinates, and abbreviations. The third set (Set C) consisted of the entire UCM herpetological database (about 17,300 records), which was used to test ease of use with large datasets, and to provide quantitative comparison between automated and manual georeferencing. This large set of data was not improved or normalized in any way. The intention was to use these record sets to compare the efficiency, consistency and precision offered by the methods considered. It is important to point out that all of these methods and tools are constantly evolving and improving, and that this discussion relates only to their status at the time we ran these georeferencing trials.

GEOLocate is a georeferencing tool developed at Tulane University (http://www.museum.tulane.edu/geolocate/demo.aspx) (Table 2). It is generally intuitive and handles both large and small datasets with relative ease and accuracy. This desktop software package combines several components of the georeferencing process into one product. It provides both automated batch georeferencing, and a user-friendly United States map interface to view and correct georeferenced points. In addition to generating Latitude/Longitude coordinates for localities, it classifies each record as having high, medium, or low precision, which is its measure of locality error. GEOLocate includes a place names database with more data than the GNIS, and is capable of georeferencing most PLSS coordinates (without _ sections), street addresses, and single and double offsets. It will return multiple results for some localities and allow the user to select the preferred result, although it only plots one of the points on its map. The map interface draws quickly and includes zoom tools, but lacks the detail necessary to completely replace USGS topographic quadrangle maps. GEOLocate does not include the measuring and route tools provided by some topographic software packages which are so useful for georeferencing. However , one can click and drag with the shift key to pan the map and measure direction and distance for measuring offsets. The GEOLocate map lacks contour lines and a PLSS or other survey grid, and does not include most feature names, which limits geovalidation. If used in concert with topographic software, this problem can be alleviated. Localities plotted on the GEOLocate map can be selected and dragged to a new location, allowing easy modification of georeferencing coordinates to match locality descriptions more exactly. Locality points can also be created manually. GEOLocate does not process elevation and highway names and numbers, and has difficulty georeferencing along linear features such as rivers and streams.

Of the 83 records in Set A, 78 were processed and returned (27 high precision, 5 medium, and 46 low). All 35 records from Set B were processed, but with low precision. More complex locality

Table 2. Summary of features and capabilities of automated georeferencing tools (Y=yes, N=no). Time units are minutes unless otherwise stated.

Feature/Capability	Biogeomancer	Georeferencing Calculator	GEOLocate	ArcView Georeferencing Extension
Format	Online	Online	Software download	Software download
Georeferencing Tool	Y	N	Y	Y
Error Estimate	N	Y	Y	Y
Error Expression	N	Point-Radius Method: Numerical error in miles	low, medium, or high precision	polygon (visual), and "span" value
Batch Processing	Y (small batches)	N	Y	N
Single Offsets	Y	Y	Y	Y
Multiple Offsets	N	N	Y	N
Abbreviations	N	N	some	N
PLSS	N	N	Y (no _ sec)	N
Place Names Database	Y	N	Y	Y (user supplied)
Street Addresses	N	N	Y	N
Highways	N	N	N	N
Elevation	N	N	N	N
Multiple Bearings (SSE, NNW, etc.)	Y	N	N	N
Text Descriptions	N	N	Y	N
Parse/Ignore Capability	N	Y	N	Y
Uses Map Interface	N	N	Y	Y
Average correct on initial georeferencing:	34%	N/A	48%	25%
Total average time per record:	6.25	5 (including manual georeferencing)	4.1	7
Time: Place Name/Coverages	0	0	0	1
Time: Database Preparation	1 (due to smaller batches)	0.5	0.5	0.5
Time: Georeferencing (coordinates)	0.25	3 (manual georeferencing)	0.1	2.5 (essentially manual georeferencing)
Time: Georeferencing (error)	2 (necessary to regraph coordinates)	0.5	2 (necessary to regraph coordinates on topo map in order to find extents)	1
Time: Quality Checking	3 (many incorrect records returned)	1	1.5 (some incorrect records returned)	2 (very difficult to check because of opaque error shapefiles)

descriptions were generally assigned low precision, which is potentially problematic, because with computer-assisted manual georeferencing, precision can be increased with the additional information provided in such complex descriptions. The 17,326 records in Set C took much longer to process than Set A or B (about 3_ hours). However, this record set showed the true power of GEOLocate by returning 16,348 of 17,326 records (8,295 high precision, 1,393 medium, and 6,654 low). Comparing the same data set, 8,379 of these records were georeferenced within one kilometer of the point chosen using the MaPSTeDI method. We were particularly impressed that 4,426 We found GEOLocate to be a useful and relatively sophisticated georeferencing tool. It correctly plotted localities that other tools could not, including the example "3 miles south of Potato Butte, Cottonwood Creek, Las Animas County, Colorado," where only a small stretch of Cottonwood Creek occurs within Las Animas County. It was able to handle formatting inconsistencies as well, although word order and some abbreviations presented problems. For example, "evans, mount," was not recognized as Mount Evans. Formatting the data prior to processing prevented most problems, but as with all methods currently available, it is also necessary to review the georeferenced results in order to identify and correct misgeoreferenced records, especially those with more complex or confusing locality descriptions.

Biogeomancer is an online georeferencing tool developed by Reed Beaman at the University of Kansas (http://www.biogeomancer.org/) (Beaman and Conn, 2003) and Yale University. Using this tool, records can be submitted and georeferenced individually or batch processed. It took about 30 minutes to prepare our first two sets of 118 total records for batch processing. This time included only the selection of required fields, ordering them as needed for submittal, and converting to delimited text format. It took only seconds for Biogeomancer to return a result set, but many records were not returned from our initial submission using both comma and tab delimited text. Only 57 of the 83 records (68%) in Set A were processed and returned. In set B, 15 out of 35 records (42%) were processed and returned. After spending another 35 minutes removing spaces, commas and parentheses from each record field, the results were slightly improved.

The data were then formatted so that the components of the locality description (coordinates, offsets and other text descriptions) were ordered uniformly, and spelled consistently, which took an additional 40 minutes. The re-formatted data were submitted once more, and a much higher number of records were processed and returned. Of the 83 in Set A, 70 (91%) were returned. Of the 35 in Set B, 32 (91%) were returned.

The results were then evaluated to determine how many of these records were georeferenced correctly. For our purposes, "correct" meant how many georeferenced records had geographic coordinates that were acceptably similar to the results of hand georeferencing as defined above, and what components of each locality record were not correctly processed by Biogeomancer and the other three tools evaluated below. Of the 83 records in Set A, 24 were correct (29%), and of the 35 in Set B, 16 were correct (45%).

Many of the correctly georeferenced records contained simple locality descriptions such

as a single offset locality. The majority of locality records in our sample set, however, were more complex. These typically contained some combination of multiple offsets, PLSS coordinates, text descriptions, obscure or unknown features, and/or geographic inconsistencies and contradictions. Single offsets with 3rd order bearings (SSW, ENE, etc.) were processed correctly, but certain common abbreviations such as ft., mi., and km. were not recognized. Biogeomancer has its own place names database which contains place names not found in the USGS GNIS database, but certain small features such as most ranches, for example, cannot be located with either. Certain geomorphological terms, such as "bluff," for example, were also not recognized.

When we attempted to evaluate Biogeomancer using Set C, but we were unable to process any data. Understandably, a web interface such as Biogeomancer was not able to process 17,000 records at once (GEOLocate took more than 3 hours to georeference the same set!), but it was also unable at the time of the experiments to process the dataset when submitted in smaller subsets (200 or 300 records at a time). Continuing upgrades to the Biogeomancer tool happening simultaneously with our experiments likely explain why the tool was unable to process our larger datasets. Biogeomancer also cannot process locality descriptions based on street addresses and roads, and does not process PLSS survey coordinates.

In conclusion, Biogeomancer works well for small sets of localities consisting of simple offsets, but the effort required preparing the locality data for processing and checking the results may not be less time consuming than computer-assisted hand georeferencing. Also, Biogeomancer does not produce an error or uncertainty value. Biogeomancer continues to grow and develop as a web service and we anticipate that future work will lead to significant performance improvements. This tool's current features and capabilities are summarized in Table 2.

The ArcView Georeferencing Extension was developed at the California Academy of Sciences (http://www.calacademy.org/research/informatics/georef/). Unlike the tools reviewed thus far, this extension turns an existing software package, ESRI ArcView, into a georeferencing tool that plots localities and produces shape files in ArcView format as a means of visualizing the locality error in a GIS (Table 2). The installation and operation instructions are directions are well written, although a working knowledge of ArcView software is extremely helpful. In addition to having ArcView software installed, the user must provide base map coverages, a place names database, and the data to be georeferenced. Depending upon the scale of the digital maps used and the area encompassed by the georeferencing project, a lot of image memory may also be necessary. The parse function permits the user to separate and prioritize words in the locality description for use in georeferencing the locality, and georeferencing takes place one record at a time. The processed record includes x, y coordinates, a place name index number (locality number), logname (name of person who submitted record), the date of georeferencing, and a shape file for each record.

Once a locality point has been georeferenced and a shape file created, the point can be moved and the shape manipulated. With numerous localities in close proximity, the view can become confusing because the locality shape files, which are defaulted to opaque, are overlapping. The shape files themselves are circles of different sizes, depending on the content of the locality description. These shapes estimate the size of the error associated with the locality description visually, but are difficult to compare to one another quantitatively because so many potential sizes and shapes can exist (although the "span" value is an expression of the largest distance within a locality polygon). On the other hand, the shape file error is an excellent way to represent linear features without having to include a circular

region when the locality occurs along a river or highway with no offset. Nevertheless, individual localities, no matter what shape, are difficult to identify on the basemap because they aren't labeled, although they can be individually highlighted using the ArcView select function. The shape files can be exported, but we were unable to export them to different themes, and we could only work within one UTM zone per project file.

We were unable to completely process Set A or B of our experimental data using the ArcView Georeferencing extension. Because it does not have a batch georeferencing function, Set C could not be tested.

In summary, the ArcView georeferencing extension is capable of processing simple offsets, but more complex locality descriptions are problematic, even after the data sets were formatted for consistency, word order, and abbreviations removed. In record by record tests, we found that the ArcView Georeferencing extension took slightly longer per record than manual georeferencing, primarily because of the time required to check the results. The cost of ArcView software might also be prohibitive. Because of its ability to represent localities and their associated error as shapefiles, however, the extension is a useful and innovative visualization tool and a large step towards efficient handling of shapefiles in georeferencing.

Developed at the University of California at Berkeley, the Mammal Information Network (MANIS) Georeferencing Calculator is an online tool which generates an estimate of error for locality data, but is not designed explicitly to georeference (http://elib.cs.berkeley.edu/manis/gc.html). In addition to processing georeferenced data (Latitude/Longitude only), its main function is to generate an error for ungeoreferenced localities one record at a time (no batch function). It requires the user to input a geographic "maximum extent" for features such as offsets to which the locality is referenced, and to estimate "distance precision." The extent is the maximum distance across the feature from one end to another. The results obtained using the Georeferencing Calculator appear to be largely based on the extent value entered, but it also takes into account all aspects of error such as datum and map error and is designed to prevent false precision (Wiecorzek et al, in press). For example, in the locality description "Bluffs, 3 miles north of Colorado Springs," one needs to enter the extent of Colorado Springs, which is about 4 miles across in the north part of Colorado Springs. The automated result returned from the Georeferencing Calculator is 4.001 (miles), which is the diameter of a circular region around the locality point which represents the Table 2. With the example "7 miles north Hudson, Colorado," we entered an extent of 0.55 miles, and obtained an identical resulting error of 0.55 miles.

Submitting each locality record, entering an extent, and obtaining an automated error from the Georeferencing Calculator took approximately 5 minutes for each of these localities, and slightly less time to georeference and obtain an error using the MaPSTeDI method. In our experimental comparisons, the Georeferencing Calculator also produced slightly larger geographic error. In conclusion, the Georeferencing Calculator is a useful tool, and represents an important contribution to the georeferencing philosophy because of the recognition of the importance of associating an error with every retrospectively georeferenced locality record.

In conclusion, the preparation of data prior to processing with any automated georeferencing tool is at this time absolutely essential. Preparation includes the removal of abbreviations, elimination or correction of unnecessary and confusing data, ordering the locality information consistently between records, and formatting as required by each georeferencing tool. All automated georeferencing tools are rapidly increasing in sophistication. Based on current and future technologies, however, a substantial post-processing quality checking effort is

required in order to correct errors generated by the automated georeferencing process. Such georeferencing errors commonly result from more complex locality descriptions. Although batch processing of records containing simple locality descriptions such as single offsets may be time effective, computer-assisted manual georeferencing is equally efficient when the number of records is low (below approximately 500). This is primarily because the latter saves time by obviating the need for exhaustive data preparation, much of which can be done as georeferencing takes place. As well, manual georeferencing also results in less incorrectly georeferenced localities, which saves substantial amounts of time in quality checking.

Of the four tools tested, only GEOlocate was able to process Set C (>17,000 records), and it was only one of three georeferencing tools that produced coordinates which were as precise and accurate as those produced by computer-assisted manual georeferencing for approximately 50% of records processed. None of the four tools evaluated contain all of the features needed for georeferencing. Although GEOlocate had the most success at automated georeferencing, it does not assign quantitative locality error values, and the most successful error generation tool did not include a georeferencing tool (MANIS Georeferencing Calculator).

There is no question that automated georeferencing tools process data much more rapidly than computer-assisted manual georeferencing methods, and that the gains in speed accrue as database sizes increase. For the largest dataset we used (set C), GEOlocate processed approximately 82 records/minute, while the MaPSTeDI method takes approximately 5 minutes/record. As discussed above, however, significant amounts of time are required to prepare (normalize) any data set for automated processing, and this is essential to obtain useable results. To normalize and prepare Sets A and B for use with Biogeomancer took 90 minutes and produced 40 “correct” records. Because most automated tools do not provide both coordinates and acceptable precision markings, time would also need to be taken to assign precision markings to each. Finally, all records would need to be checked to verify that the georeferencing was done correctly. This final validation step is potentially as time consuming as any other step, although map features such as those used by GEOLocate can decrease this validation time considerably.

Most of the tasks described above – database normalization, the assignment of spatial coordinates, and error determination - can be performed simultaneously when manually georeferencing. In addition, manual georeferencing also generally results in less incorrectly georeferenced localities, which saves substantial amounts of time in quality checking. All this boils down to an average of 5 minutes per record, while automated tools vary more in their average time per record. Time estimates broken down by each step of georeferencing for the four automated methods have been listed in Table 2. We stress that these are estimates, and obviously results will vary from dataset to dataset due to size, complexity, etc.... In the short term, one obvious solution to maximize efficiency for larger dataset is to separate records into simple and more complex locality descriptions using classification schemes like Table 1. The simple records are likely good ones for automated tools, while more complex records are at this juncture better coded using a computer-assisted manual method.

Future directions in georeferencing technologies

The time is right to integrate knowledge derived from the multiple manual and automated tool developments for georeferencing. Recently funded projects (GBIFs’ DIGIT projects, MANIS, HERPNET, MaPSTeDI, ORNIS, INRAM) for georeferencing natural history museum collections have generated rules for manual georeferencing but are slow. Automated

tools (Biogeomancer, GEOLocate, Manis Georeferencing Calculator, ADEPT, CAS ArcView Extension) for desktop and the Web have tried to incorporate some of the simpler rules for georeferencing but are not sophisticated enough yet. With recent technological advances that support distributed computing architectures (SOAP, XML, WSDL, and UDDI), there is an opportunity to work collaboratively to build a next generation georeferencing tool which incorporates lessons learned to date, facilitates data standardization, and lowers the overall cost for georeferencing natural history museum collections data.

One reason for presenting our georeferencing method is that ultimately, manual georeferencing methods need to be incorporated into the next-generation automated tools. We have argued that the current manual georeferencing methods are more reliable and more accurate than current automated tools. The complexities of natural-language information that make up locality descriptions are much more effectively processed by humans following flexible but specific rules than computers following vague and inflexible algorithms. Existing manual protocols (of which there are three for natural history collections – MaNIS, MaPSTeDI, and INRAM) will need to be reconciled and a consensus set of types for localities and rules associated with georeferencing those types generated. This consensus will serve as the backbone for more sophisticated semi-or fully automated next-generation tools. Once automated tools incorporate natural-language processing rules and better validation methods, they have the potential to provide very fast, accurate and precise georeferencing.

We envision that a collaboratively built next generation georeferencing tool would be provided as a web-based solution. Widely accessible and continually updated, this web-based solution would provide those involved in georeferencing natural history data with access to a gazetteer portal, spatially referenced basemap data, data input forms, automated accuracy and precision estimation, and data export capabilities. The search for already georeferenced localities would access a gazetteer portal consisting of all known georeferenced localities recorded to date from web based services such as the USGS GNIS, ADL Gazetteer, and data sets compiled from such projects as DIGIT, MANIS and MaPSTeDI. Access to spatially referenced basemap data would come from services such as TerraServer USA, National Geographic's Topo! Map service for U.S. based localities, and NIMA's VMAP for non-U.S. based localities. Both manual and automated georeferencing tools would be provided in which a user could upload existing digital records and then iteratively or batch process the records. During the georeferencing process, automated accuracy estimators would assign an uncertainty associated with each georeferenced record. At the conclusion of the georeferencing process, the results would be available to download in various export options such as tab-delimited text files, EXCEL spreadsheets, or shapefiles.

Above all, any next-generation tool would not eliminate the human aspect from georeferencing, but instead provide the user everything required to make a quick and intelligent decision about georeferencing a specimen. Georeferencing remains perhaps the essential task that needs to be completed across all museums in the world before we can unlock the full utility of museum collections for biodiversity analysis. We believe community based tool building based on best practices documents such as this one represent the best opportunity forward.

Acknowledgements

We would like to thank all the undergraduate workers and graduate assistants who helped with the georeferencing during the MaPSTeDI grant. Georeferencing is an iterative process and we learned as much as we did by exploring the full parameter space of types of georeferenceable records through our student helpers. John Wieczorek at the Museum of Vertebrate Zoology, University of California at Berkeley,

developed protocols for georeferencing almost simultaneously with MaPSTeDI, and has been willing to share his similar approaches and discuss shared problems in georeferencing. We appreciate the opportunity to participate in an NSF and GBIF sponsored workshop at Yale, run by John Wieczorek and Reed Beaman, and our discussion of the future of georeferencing tools reflects some of the excellent discussions at that meeting. Support provided by NSF grant DBI-0110133 to PI R. P. Guralnick is gratefully acknowledged.

References Cited

Alberch, P., 1993. Museums, collections and biodiversity inventories: Trends in Ecology and Evolution 8:372-375

Baker, R.J., Phillips, C.J., Bradley, R.D., Burns, J.M., Cooke, D., Edson, G.F., Haragan, D.R., Jones, C., Monk, R.R., Montford, J.T., Schmidly, D.J., and Parker, N.C., 1998. Bioinformatics, museums and society: Integrating biological data for knowledge-based decisions: Occasional Papers, Museum of Texas Tech University 187:1-4.

Beaman, R., and Conn, B. 2003. Automated geoparsing and georeferencing of Malesian collection locality data: Telopea 10(1):43-52.

Brisby, F. 2000. The Quiet Revolution: Biodiversity Informatics and the Internet. Science 289: 2309-2312.

Coterill, F.P.D. 1995. Systematics, biological knowledge and environmental conservation: Biodiversity and Conservation 4:183-205.

Fischman, J. 1996. Bioinformatics: Working the web with a virtual lab and some java: Science 273:591.

Grace, J.B., 1997. Bioinformatics: Mathematical challenges and ecology: Science 275:861.

Krishtalka, L., and Humphrey, P.S. 2000. Can natural history museums capture the future: Bioscience 50(7):611-617.

D. Neufeld and Guralnick, R. P. In review. Research Challenges in Using Distributed GIS Services to Support Biodiversity Visualization and Analysis. International Journal of Geographical Information Science.

Parker, N.C., Bradely, R.D., Burns, J.M., Edson, G.F., Haragan, D.R., Jones, C., Monk, R.R., Montford, J.T,., Phillips, C.J., Phillips, D.J., Schmidly, D.J., and Baker, R.J., 1998. Bioinformatics: A multidisciplinary approach for the life sciences. Occasional Papers, Museum of Texas Tech University 186:1-8.

Stein, B. and Wieczorek, J. In press. Mammals of the World: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics.

Wiecorzek, J.W., Guo, Q., and Hijmans, R.J. in review. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty: International Journal of Geographical Information Science.

Wilson, E. O. 1992. The Diversity of Life. (Cambridge, UK, Belknap Press).

Internet Sites

ADEPT, 2001. ADEPT Educational Collections, ADEPT. URL http://piru.alexandria.ucsb.edu/

Blum, Stan, 2001. Georeferencing Natural History Locations at the California Academy of Sciences, California Academy of Sciences. URL http://www.calacademy.org/research/informatics/georef/

Beaman, Reed, (2002). Biogeomancer, University of Kansas. URL http://www.biogeomancer.org/

Tulane University, (2003). GEOLocate Georeferencing Software for Natural History Collections, Tulane University. URL

http://www.museum.tulane.edu/geolocate/demo.aspx

Spencer, Carol, (2003). HERPNET, University of California at Berkeley. URL http://www.herpnet.org/

Institue of Natural Resource Analysis and Management, 2002. Inram: Institue of Natural Resource Analysis and Management INRAM. URL http://www.inram.org/

Wieczorek, John, 2001. MANIS Georeferencing Calculator, University of California at Berkeley. URL http://elib.cs.berkeley.edu/manis/gc.html

Mapquest.com, 2004. Mapquest: Home, Mapquest.com. URL http://www.mapquest.com/

Guralnick, Robert, et al, 2003. Mountain and Plains Spati-Temporal Database-Informatics Initiative (MaPSTeDI), University of Colorado at Boulder. URL http://mapstedi.colorado.edu

NOAA. FGCS – Selective Availability Removal, National Oceanic and Atmospheric Administration. URL http://www.ngs.noaa.gov/FGCS/info/sans_SA/

SSIG, 2003. ORNIS, SSIG. URL http://www.kbinirsnb.be/cb/ornis/

Santa Cruz Public Library System, (2000). Highway Numbering System, URL http://www.santacruzpl.org/readyref/files/g-l/hiwaynos.shtml

TopoZone, 2004. TopoZone – The Web's Topgraphic Source and More!, URL http://www.topozone.com/

USGS, 2004. Geographic Names Information System (GNIS), United States Geological Survey, URL http://geonames.usgs.gov

Appendix 1
EXAMPLES AND TREATMENTS FOR VAGUE LOCALITY DESCRIPTIONS.

Example: Near Boulder Falls, Boulder County, Colorado.
Treatment: Unfortunately, “near” is an all to common descriptor in older records. These should be assigned the coordinates of the primary feature named in the locality, in this case, Boulder Falls. An appropriate confidence value should be selected that would encompass the maximum estimated area of error. The area of error can be difficult to determine, because “near” is a subjective judgement which means different things to different workers.

Example: 5 miles from Boulder Falls, Boulder County, Colorado.
Treatment: Offsets lacking a direction are often the result of an accidental omission by the collector at the time the locality was recorded. These should be assigned the coordinates of the primary feature named in the locality, in this case, Boulder Falls. An appropriate confidence value should be selected that would encompass the maximum estimated area of error, in this case, 5 miles. Because offsets lacking a direction may also be data entry errors, the original collections catalogue and field notes should also be consulted.

Example: North of Boulder Falls, Boulder County, Colorado.
Treatment: In cases when locality offsets are lacking a distance from the referenced feature, the locality should be assigned the coordinates of the northern boundary of the feature. If the feature is too small to have a northern boundary, the point should be placed on the feature. In either case, a confidence value should be assigned which encompasses the maximum estimated area of error. Because offsets lacking a distance may also be data entry errors, the original collections catalogue and field notes should also be consulted.

Example: North of Boulder Falls on CO SH-119, Boulder County, Colorado.
Treatment. This is an example of a locality description containing contradictory information, because State Highway 119 is actually south of Boulder Falls. Because the contradiction may simply be the result of a data entry error, the collections catalogue and field notes should be consulted and compared to the record if possible. In all such cases, it may be necessary to consult older maps and gazetteers in order to determine if either feature may have changed in location since the locality was first documented. If it is determined that the contradiction in the locality description was not the result of either data entry or a change in feature location, it may be necessary to use Boulder Falls as the locality point, and encompass State Highway 119 with the confidence value. In this case though, the locality description indicates that the point is actually “on” State Highway 119. With this information, the georeferencer can reasonably deduce that the description should have read south instead of north, plot the locality point appropriately, and document the rationale for the decision in the record progress field.

Example: Boulder Creek, Boulder County, Colorado.
Treatment: Localities listing only a linear feature such as a road or river are difficult or even impossible to plot with an acceptable level of precision because linear features are often quite long. With no other information, the georeferencer should assign coordinates for a point midway along the feature with an appropriate confidence value. In this case, the point would be exactly halfway along Boulder Creek within Boulder County, Colorado. More precise positioning of linear feature localities is possible if an intersecting feature such as a town, highway or elevation is included in the locality description.

Example: Boulder County, Colorado.
Treatment: Localities described to only the level of county, state or other large geographic area cannot be georeferenced to a level useful for most biodiversity-related inquiries. The georeferencer should leave the coordinates field blank, and assign the appropriate confidence value. Despite their imprecision, note that these records can be included and displayed in a GIS using county, state or other regional polygon coverages.

APPENDIX 2.
EXAMPLES AND APPLICATION OF MAPSTEDI CONFIDENCE INTERVALS

Example, CV 1: Confluence of North and Middle Boulder Creek, 465408E, 4428096N, Boulder County, Colorado.
Treatment: Localities receiving a CV of 1 are typically the simplest to georeference because UTM or Lat/Long coordinates are provided in the locality description, making the highest degree of precision possible. It is important, however, to verify that the coordinates are correct by comparing them with the other information in the locality description. If the coordinates are incorrect, it may be necessary to assign a less precise CV.

Example, CV 2: Boulder Falls, 465500 mE, 4428500 mN, Boulder County, Colorado
Treatment: A CV of 2 is assigned in several instances. It is used if the field collector documented the locality using a datum different from that used in the georeferencing process (NAD27 in MaPSTeDI). This is then corrected by converting the original coordinates using the new datum, which is readily accomplished using most topographic software programs. A CV of two can also be used if, for example, the field worker rounded off the UTM coordinates for the locality (see example above). If the coordinates appear to be rounded off, they should be checked against the rest of the locality description, corrected if necessary, and a CV of 2 assigned.

Example, CV 3: Boulder Canyon, PM 6 T1N R72W Sec36, SE_, NW_, SE_, Boulder County, Colorado.
Treatment: If the locality description includes Public Land Survey System coordinates, a CV of 3 is used. As with the examples above, these coordinates should be verified by comparing then with other information in the locality description, if it exists. If they do not match, it may be necessary to assign a less precise CV.

Example, CV's 4, 5, and 6: City of Boulder, Colorado.
Treatment: These CV's are assigned to localities lacking UTM, Lat/Long, or PLSS coordinates, and have a larger area of error which is represented by a circular region surrounding the locality point selected by the georeferencer. This region encloses all other possible locations for the point based on the information provided in the locality description, depending upon which CV is selected. A CV of 4 encloses all possible points within a 1 km radius of the georeferenced point, a CV of 5 encompasses a 5 km radius, and a CV of 6 a 10 km radius. The 1, 5, and 10 km categories were chosen by the MaPSTeDI Project because they are easy for georeferencers to work with, and most localities were georeferenced with a CV of 4 or 5.

Using the example above, the City of Boulder, Colorado, can be enclosed within a circular area with a radius of 5.6 km. Therefore, a CV of 6 would be assigned, because the precise location of the original locality within the City of Boulder is unknown, and is probably not the exact center of the city (wherever that may have been at the time the locality was first documented). Therefore, the estimated area of error could be as large as a 5.6 km circle. The CV assigned to this locality would be 6 because the margin of error is less than 10 km but greater than 5 km. Note that in addition to localities with a margin of error of up to 1 km, a CV of 4 is assigned to more precise locality descriptions without UTM, Lat/Long, or PLSS coordinates such as "Boulder Falls, Boulder County, Colorado," for example. As with the assignment of coordinates, the rationale for the CV assignment should be documented by the georeferencer.

Assigning CV's is more difficult with more complex locality descriptions. When assigning a CV for a locality with an offset, for example, the CV must take into consideration the size reference feature for the offset. A locality description such as "2 miles east of Boulder, Colorado," would be assigned a CV of 6 because of the distance represented by the range of possibilities for the origin of the offset within Boulder. Using the present size of Boulder as reported above (5.6 km across), if the original locality was actually on the western edge of Boulder, 2 miles (3.2 km) east of that point would actually be 2.4 km within the present city limits. If the original locality was in the center of Boulder, 2 miles (3.2 km) east of that point would be only 0.2 miles (0.4 km) east of the eastern edge of Boulder. Note that using the MaPSTeDI method, the GNIS point for the center of Boulder would be used as the point for the origin of the offset, the georeferenced locality point would be plotted 2 miles (3.2 km) east of the GNIS point, a CV of 6 would be assigned because 2 miles east of any possible point within Boulder would fit within a 10 km circular area, and the rationale for the georeferencing decision and CV assignment would be documented by the georeferencer. Given more locality information, the margin of error could be greatly reduced. If the locality description was "2 miles east of Broadway in Boulder," the CV could be reduced to a 4, and the locality positioned with much greater precision.

The CV assignment must also take into account uncertainty relating to offset direction and length. In general, the georeferencer should assume that the distance and direction of an offset is accurate. However, as demonstrated above, there are localities for which it is difficult to establish the location of the offset reference feature precisely. In these cases, the area of possible error (CV) should be expanded to include all possible locations of the point.

Example, CV 7: Boulder County, Colorado.
Example, CV 7: Boulder Creek, Boulder County, Colorado.
Treatment: When the county is the most precise locality information provided, or when an area of error greater than a 10 km radius but less than the size of the county is indicated, a CV of 7 is assigned. Examples of the latter often include locality descriptions containing only a linear feature such as a river and the county of collection, with resulting large areas of uncertainty.

Example, CV 8: Eastern plains, Colorado
Example, CV 8: South Platte River north of Denver, Colorado
Treatment: This CV is used for records with little or no information more precise than state. It is also used with localities in which the county is not named, and which there is a large area of error typically resulting from vague or otherwise insufficient geographic information. The CV of 7 should not be used when the margin of error can also be described with a CV of 4 through 6. For example, the locality "Just east of Norfolk on I-25, Colorado" without a county reference could be located in either Larimer or Weld counties. Although the range of possible locations and resulting georeferenced coordinates includes parts of two counties, the area indicated by the locality description can be encompassed with a circle with a radius under 10 km, and should therefore be assigned a CV of 6 instead of a CV of 8.

Example, CV 9: Rocky Mountains
Treatment: For localities which could be located in more than one state, and those even more unfortunate records for which only the country is listed, a CV of 9 is assigned. As with the CV of 8, a CV of 9 should not be used when the margin of error can be described with a CV of 4 through 6. This could occur if the locality description included more precise geographic information that made it possible to encompass the area of error with a circle which was 10 km or less in size, as exists along borders between states.

4.3.2. Vagueness and Confidence Values
For vague locality descriptions, it is suggested that the CV be raised one level, or otherwise adjusted to address the vagueness as appropriate. Documentation of the rationale behind the decision should be attached to the georeferenced record.

Example: Near Boulder Falls, Boulder County, Colorado.
Treatment: Using MaPSTeDI protocols, the locality "Boulder Falls, Boulder County, Colorado," would typically be assigned a CV of 4. Therefore, as indicated above, "near Boulder Falls, Boulder County, Colorado," would receive a CV of 5.

Example: 5 miles from Boulder Falls, Boulder County, Colorado.
Treatment: With an offset that lacks a direction, the size of the area of potential error is doubled. For example, although the locality described here is only 5 miles (8 km) from Boulder Falls, there is no direction specified. Therefore, the area of potential area could be as much as 10 miles (16 km) cross, and a CV of 7 should be assigned.

Example: North of Boulder Falls, Boulder County, Colorado.
Treatment: For offsets with no distance, the area of potential error is impossible to determine with any precision. The MaPSTeDI treatment for such records involves evaluation on a case by case basis, and documentation of the rationale for the decision. In this case, the locality point would be plotted 1 km north of the GNIS point for the Boulder Falls, and makes the assumption that a different reference point for the locality would have been chosen if it were more than 1 km north of the waterfall. Therefore, the CV would be 4.