We have seen how it is possible to download occurrence records from a platform such as the GBIF, and how to store them on your HD for further use and reuse.
As a best practice, it is always a good idea to store both the original download, and the data you have prepared and cleaned.
Any data frame can be exported in the CSV format (comma-separated values). This format is an international standard, and is properly read by practically any database, and spreadsheet.
Values can be separated by commas, or by other tags. The most common is the TAB (^t), but one can use anything.
By default, the R command write.csv() exports in a CSV file with values separated but TAB.

In our example, we used the command:
write.csv(dl, "dulcamara_GBIF.csv")
Which save the data frame **dl** in the CSV file **dulcamara_GBIF.csv**

We can specify many other options in the command, such as a different separator, encoding, etc.

The R command *gbif()* allows to download data from the GBIF. It is included in the package *dismo*. There exist other commands in other packages, which are alternative to this one, and do practically the same things.

In our example, we used the command:
gbif("solanum", "dulcamara", download=FALSE, geo=FALSE, args = c(‘country=IT'))*
To count the occurrences, and then the command:
gbif("solanum", "dulcamara", geo=FALSE, args = c('country=IT'))*
*w*hich queries the **GBIF** for the occurrences of a taxon (***Solanum dulcamara***, the query is key insensitive), without distinction among them as far as geo-reference are concerned (for the *geo* parameter, **TRUE** excludes non geo-referenced occurrence data, while **FALSE** downloads everything), specifying that we want the occurrences from Italy (**IT**, GBIF uses two character country IDs) alone. The parameter **download** allows to decide if count (**FALSE**) or download (**TRUE**, or skipping the parameter) the occurrences.

We also store the downloaded data into a data frame (**dl**), for further usage.

The R command *gbif()* has potentially many other parameter which can be specified:

genus - character. genus name
species - character. species name. Use '*' to download the entire genus. Append '*' to the species name to get all naming variants (e.g. with and without species author name) and sub-taxa
ext - Extent object to limit the geographic extent of the records.
args - character. Additional arguments to refine the query. See query parameters in http://www.gbif.org/developer/occurrence for more details
geo - logical. If TRUE, only records that have a georeference (longitude and latitude values) will be downloaded
sp - logical. If TRUE, geo will be set to TRUE and a SpatialPointsDataFrame will be returned
removeZeros - logical. If TRUE, all records that have a latitude OR longitude of zero will be removed if geo==TRUE, or set to NA if geo==FALSE. If FALSE, only records that have a latitude AND longitude that are zero will be removed or set to NA
download - logical. If TRUE, records will be downloaded, else only the number of records will be shown
ntries - integer. How many times should the function attempt to download the data, if an invalid response is returned (perhaps because the GBIF server is very busy)
nrecs - integer. How many records to download in a single request (max is 300)?
start - integer. Record number from which to start requesting data
end - integer. Last record to request

Let's see a further example, making more data cleaning, an also some georeferencing, by using the Google API. This can be done only if your R environment has the API set.

We are:

1. Creating a data frame
2. Cleaning duplicates
3. Getting rid of non geo-referenced records
4. Getting rid of wrong geo-referenced records
5. Correcting some incorrect geo references
6. Geo-reference some non geo-referenced records
7. Resample to get rid of oversampling

Let’s switch to R again

Georeferencing

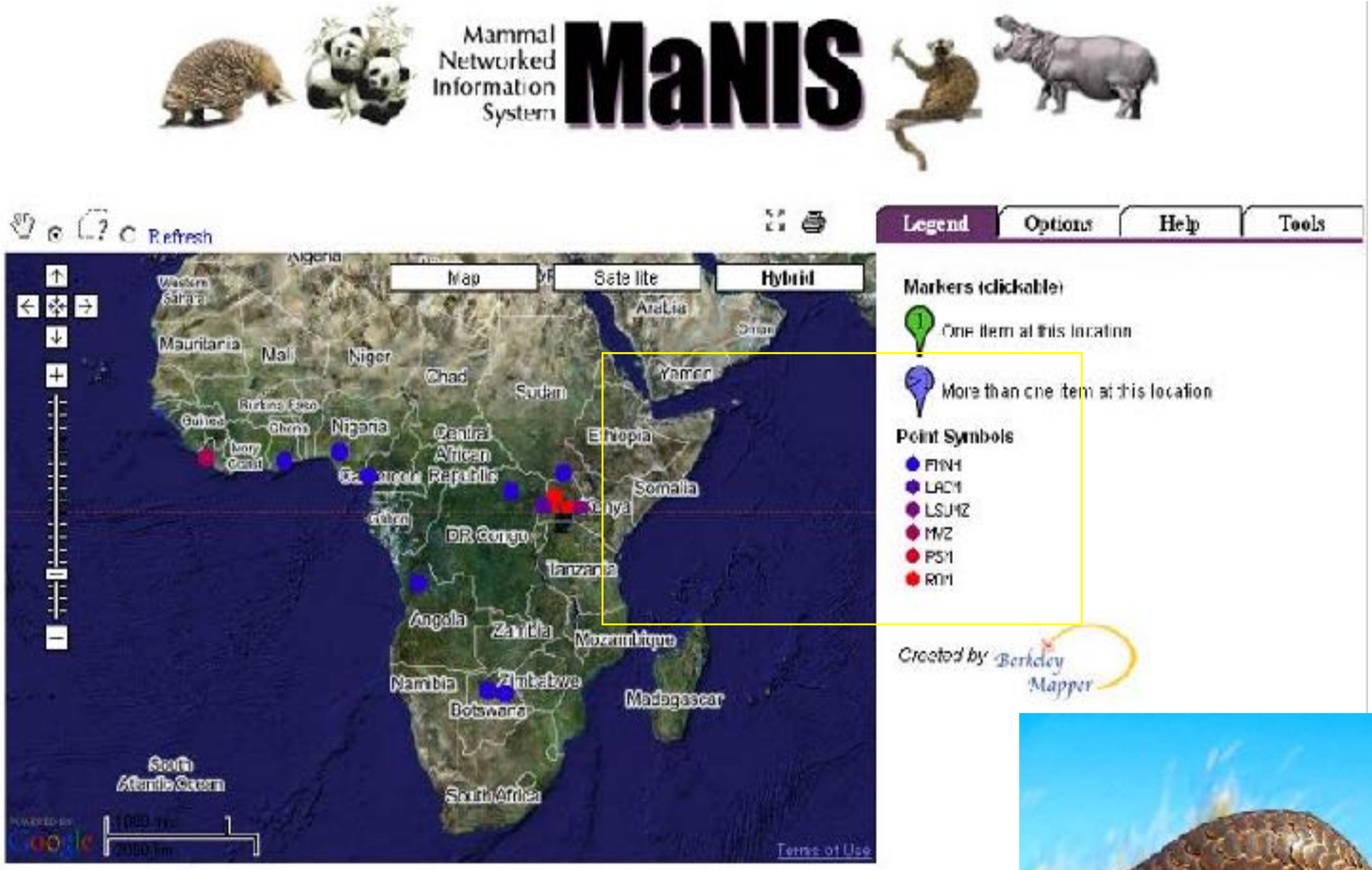
What a georeference is?

A numerical description of a place that can be mapped.

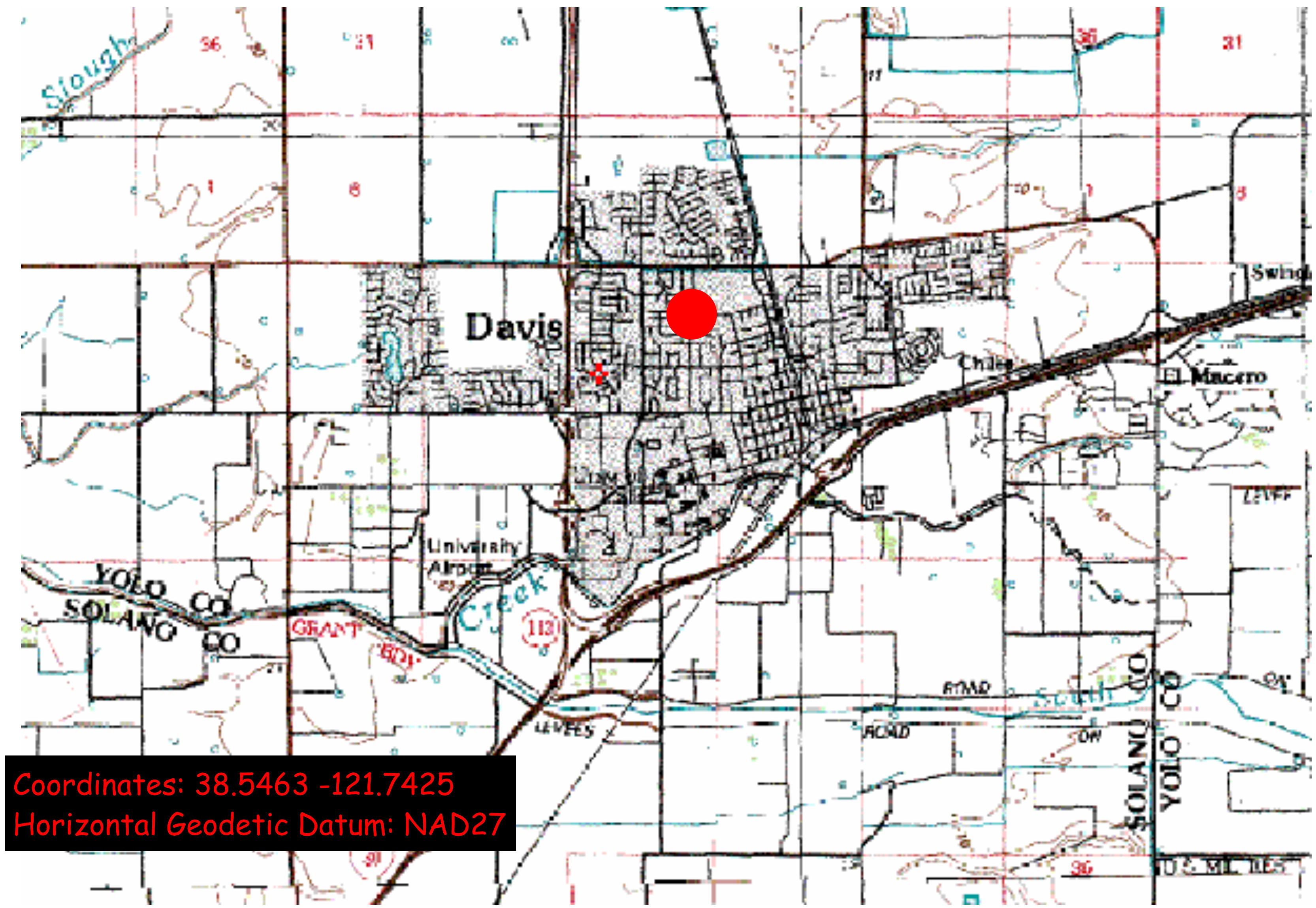
What we have:
Localities we can read

ID	Species	Locality
1	Lynx rufus	Dawson Rd. N Whitehorse
2	Pudu puda	cerca de Valdivia
3	Canis lupus	20 mi NW Duluth
4	Felis concolor	Pichi Trafúl
5	Lama alpaca	near Cuzco
6	Panthera leo	San Diego Zoo
7	Sorex lyelli	Lyell Canyon, Yosemite
8	Orcinus orca	1 mi W San Juan Island
9	Ursus arctos	Bear Flat, Haines Junction

What we want:
Localities we can map



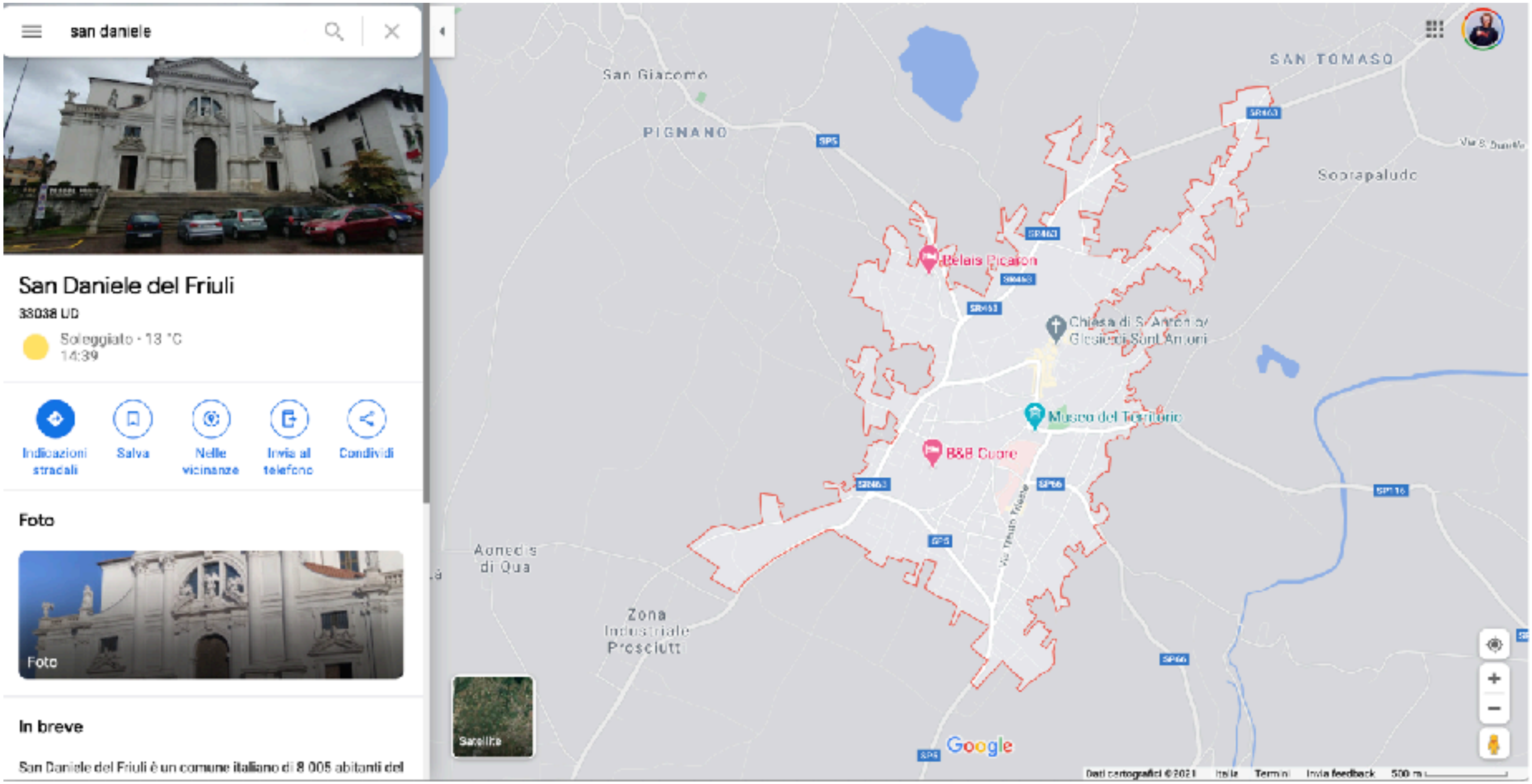
“Davis, Yolo County, California”



“point method”

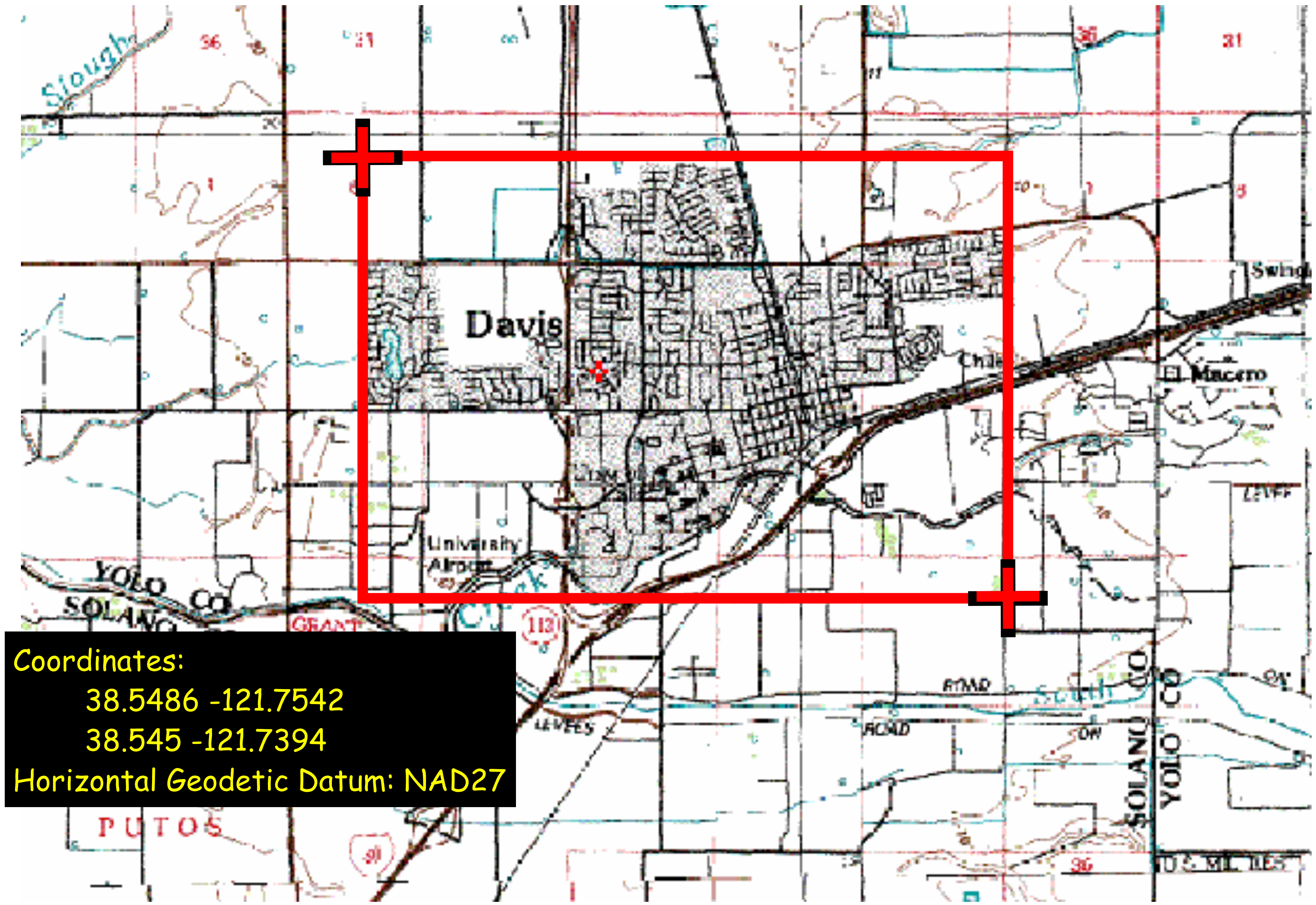
What is an acceptable georeference?

A numerical description of a place that can be mapped and that describes the spatial extent of a locality.



Every named place occupies a finite space, or ‘extent’.

“Davis, Yolo County, California”



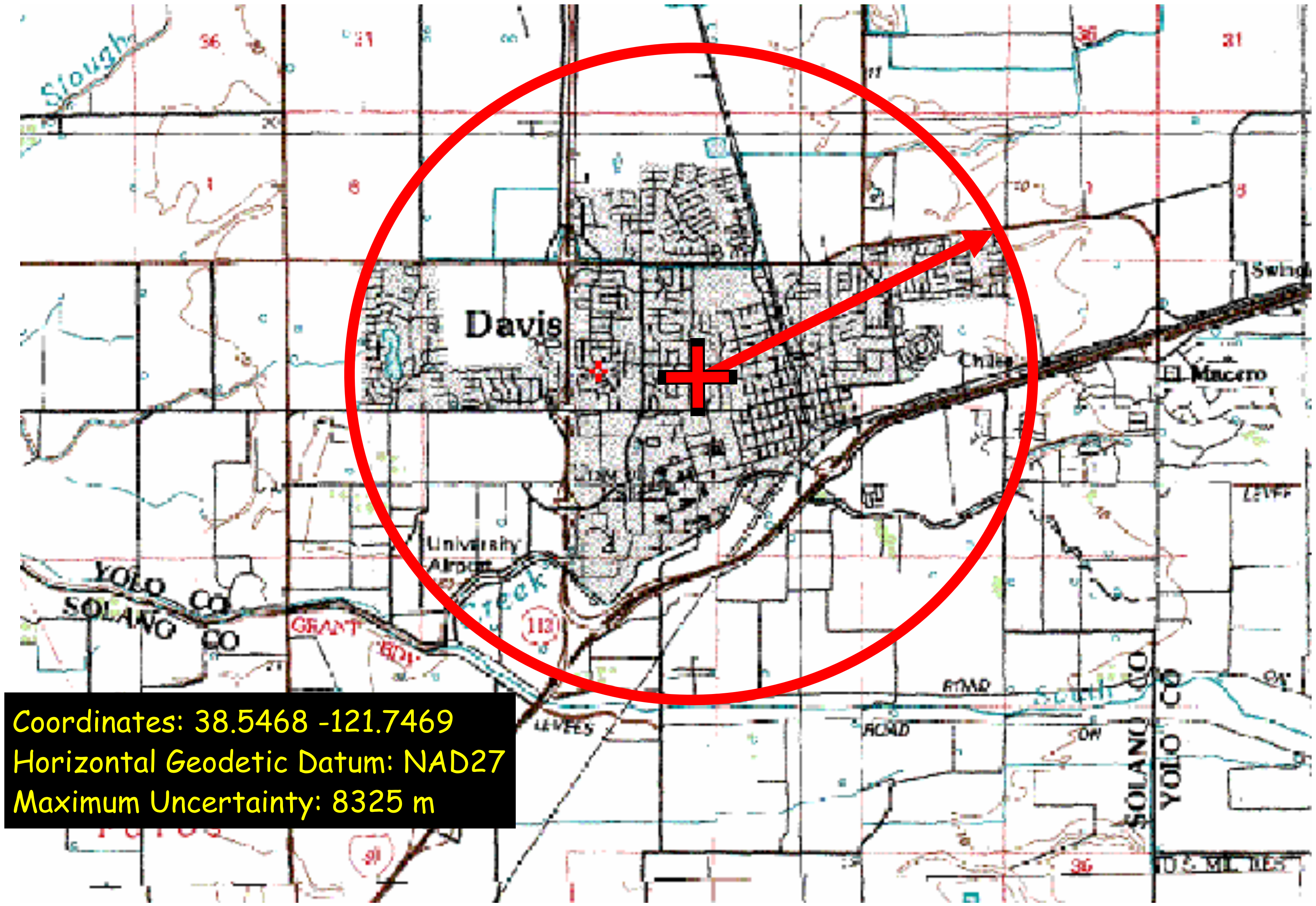
Coordinates:
38.5486 -121.7542
38.545 -121.7394
Horizontal Geodetic Datum: NAD27

“bounding-box method”

What is a high quality georeference?

A numerical description of a place that can be mapped
and that describes the spatial extent of a locality
and its associated uncertainties.

"Davis, Yolo County, California"



"point-radius method"

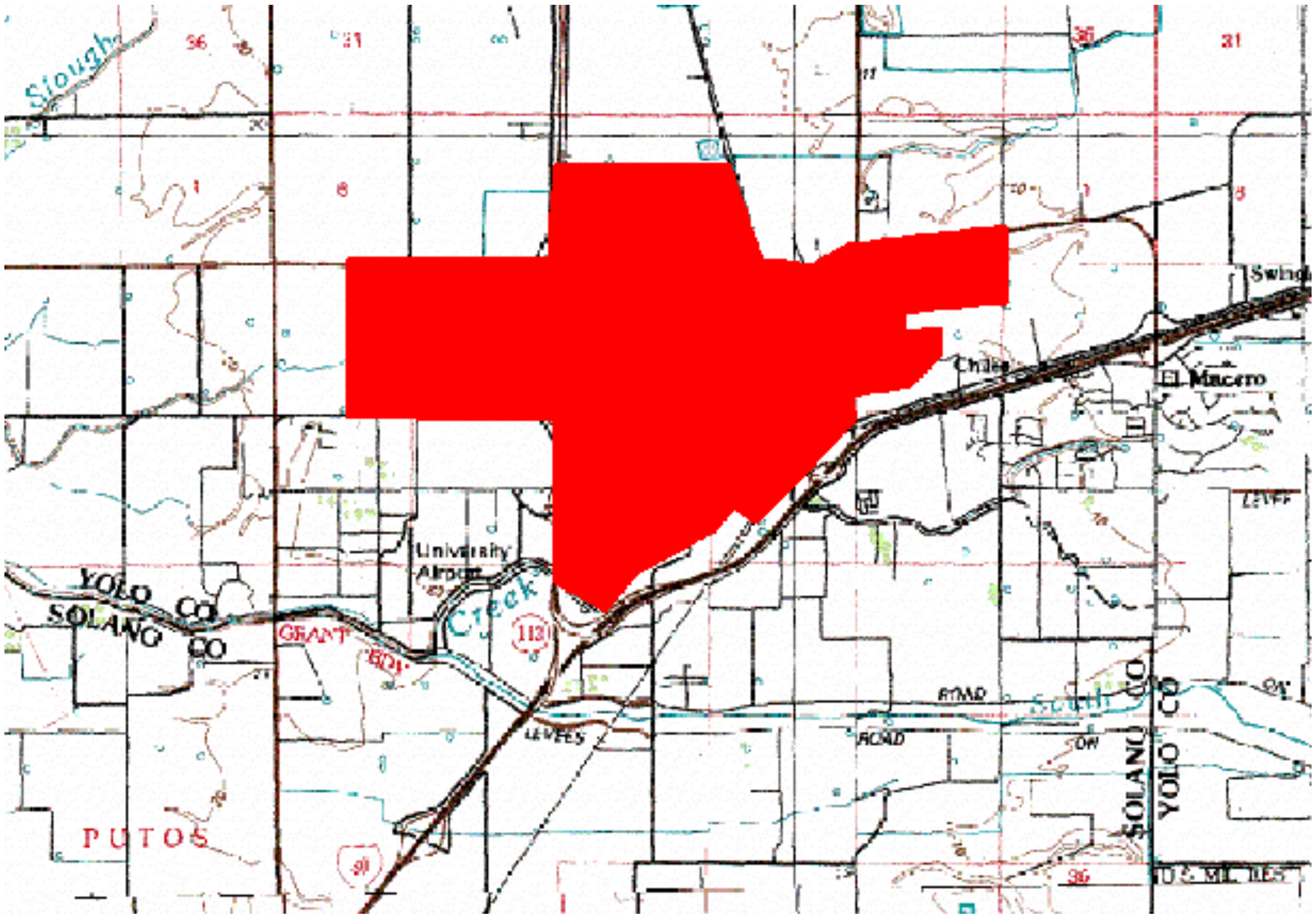
The geographic centre (that is, the midpoint of the extremes of latitude and longitude) of the named place is recommended as the location of the coordinates because it describes a point where the uncertainty due to the extent of the named place is minimized.

If the locality describes an irregular shape (for example, a winding road or river) and the geographic centre of that shape does not lie within the locality, then the point nearest the geographic centre that lies within the shape is the preferred reference for the named place and represents the point from which the extent of and offsets from that named place should be calculated.

What is a ideal georeference?

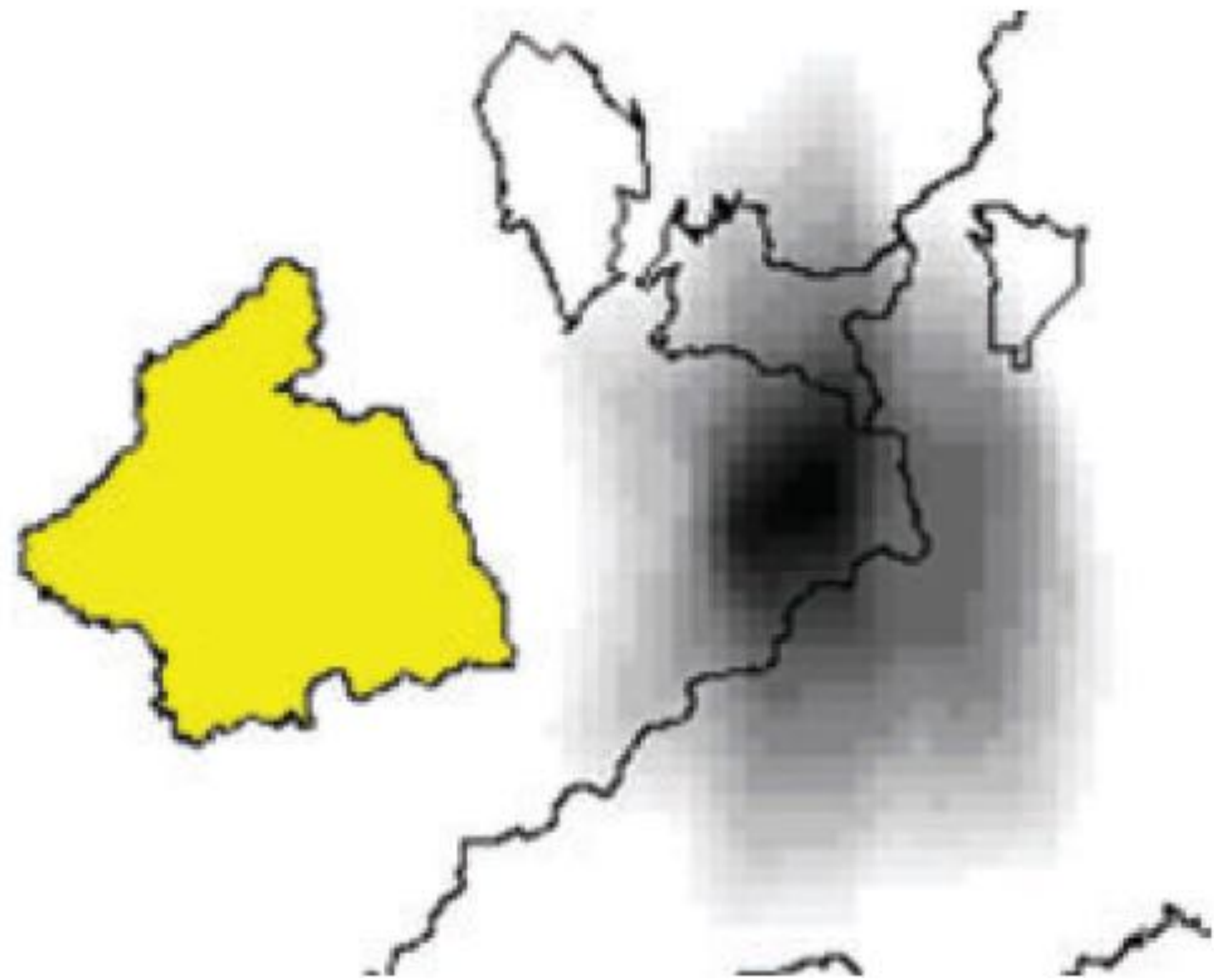
A numerical description of a place that
can be mapped
and that describes the spatial extent of
a locality
and its associated uncertainties
as well as possible.

"Davis, Yolo County, California"



"shape method"

“20 mi E Hayfork, California”



“probability method”

Method comparison

point	easy to produce
	no data quality
bounding box	simple spatial queries
	difficult quality assessment
point-radius	easy quality assessment
	difficult spatial queries
shape	accurate representation
	complex, uniform
probability	accurate representation
	complex, non-uniform

The **point-radius method** provides a practical solution for georeferencing descriptive localities that can be widely implemented, especially in communities where sophisticated GIS expertise is lacking.

By accounting for the size of the locality, the point-radius method provides a more accurate description of a locality than is possible with the point method.

By providing a single measure of the combination of uncertainties inherent in the locality description, the applicability of a locality for a given analysis can be more readily discerned than with the bounding box method.

By capturing the spatial attributes of the locality in a simple, consistent set of parameters, the point-radius method offers a solution that is practical for natural history collections without the need for spatial databases that would be necessary to store georeferences created using the shape method.

Why georeference?

Map species locations

Understand species ranges

Enable spatial analyses

Combine with other spatial data

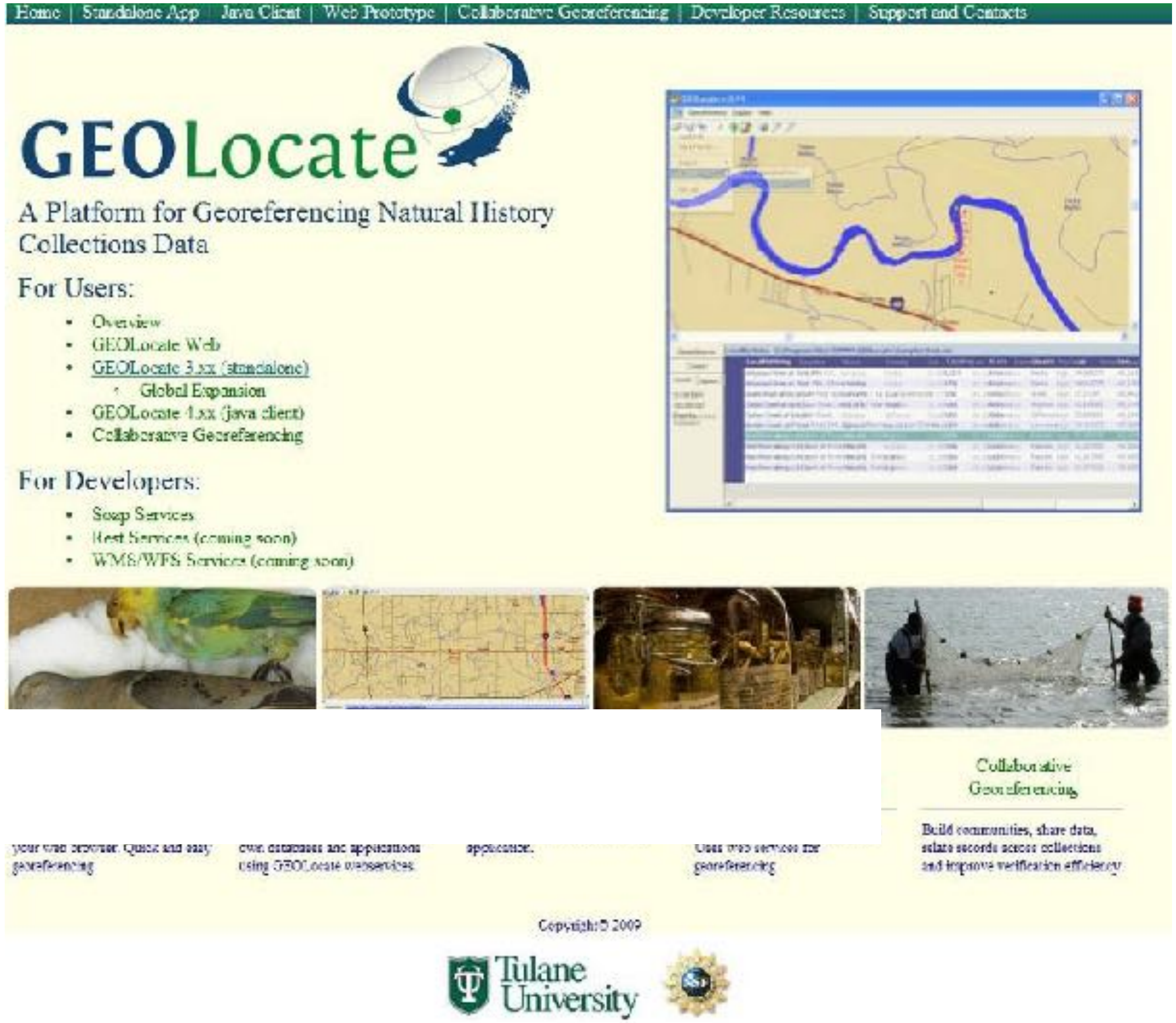
Georeferencing Sources

Specimen labels
Field notes
Literature

Gazetteers
Printed Maps
Digital Maps

Table 1. Types of locality descriptions commonly found in natural history collections.

Type	Description	Examples
1) dubious	The locality explicitly states that the information contained therein is in question.	'Isla Boca Brava?', 'presumably central Chile'
2) can not be located	Either the locality data are missing, or they contain other than locality information, or the locality cannot be distinguished from among multiple possible candidates, or the locality cannot be found with available references.	'locality not recorded', 'Bob Jones', 'lab born', 'summit', 'San Jose, Mexico'
3) demonstrably inaccurate	The locality contains irreconcilable inconsistencies.	'Sonoma County side of the Gualala River, Mendocino County'
4) coordinates	The locality consists of a point represented with coordinate information.	'42.4532 84.8429', 'UTM 553160 4077280'
5) named place	The locality consists of a reference to a geographic feature (e.g., town, cave, spring, island, reef, etc.) having a spatial extent.	'Alice Springs', 'junction of Dwight Avenue and Derby Street'
6) offset	The locality consists of an offset (usually a distance) from a named place.	'5 km outside Calgary'
7) offset along a path	The locality describes a route from a named place.	'1 km S of Missoula via Route 93" "600 m up the W Fork of Willow Creek'
8) offsets in orthogonal directions	The locality consists of a linear distance in each of two orthogonal directions from a named place.	'6 km N and 4 km W of Welna'
9) offset at a heading	The locality contains a distance in a given direction.	'50 km NE Mombasa'



http://www.geo-locate.org/default.html

RDocumentation | Search for packages, functions, etc | R Enterprise Training | R package | Leaderboard | Sign in

biogeomancer

From dismo v0.9-1 by Robert Hijmans | 95th Percentile

Georeferencing

A link to the biogeomancer georeferencing web service. See http://bg.berkeley.edu/latest/ for more information and a rich visual interface.

Keywords spatial

Usage

```
biogeomancer(country = '', adm1 = '', adm2 = '', locality = '',
             singleRecord=TRUE, progress='text')
```

Arguments

country Character. Country

adm1 Character. Name of the first-level administrative subdivision. E.g. the State in the United States and India, the Province in China and Canada

adm2 Character. Name of the second-level administrative subdivision. E.g. the county in the United States, and the district in India and China

locality Character. Locality description

singleRecord Boolean. If `TRUE`, the record with lowest uncertainty is selected when several records are returned

progress Character. Valid values are "" (no progress indicator), "text" (the default) and "windows" (on that platform only)

Value

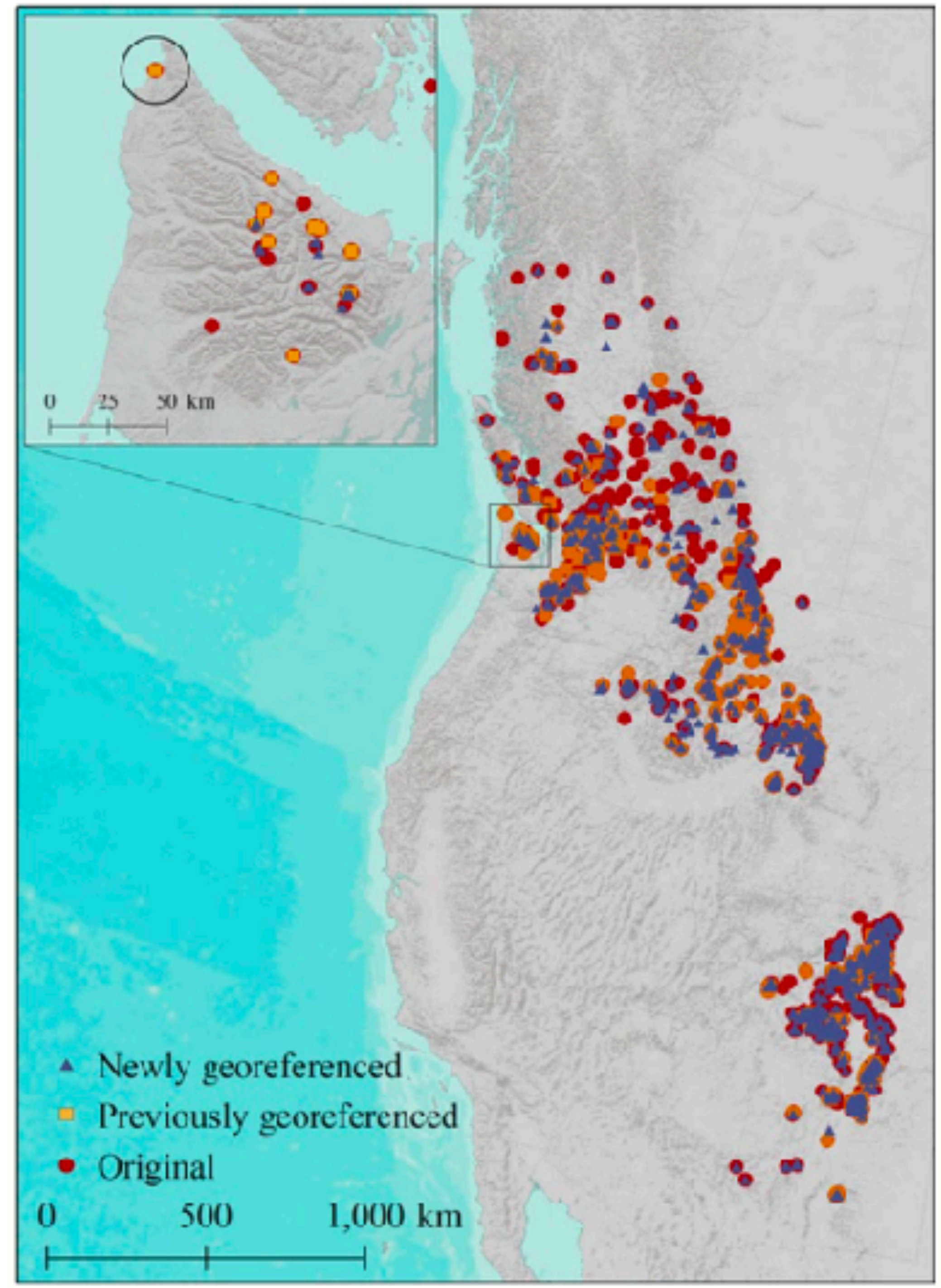
data frame with three columns: longitude, latitude, and uncertainty (see Wieczorek et al., 2004). The datum is always WGS84.

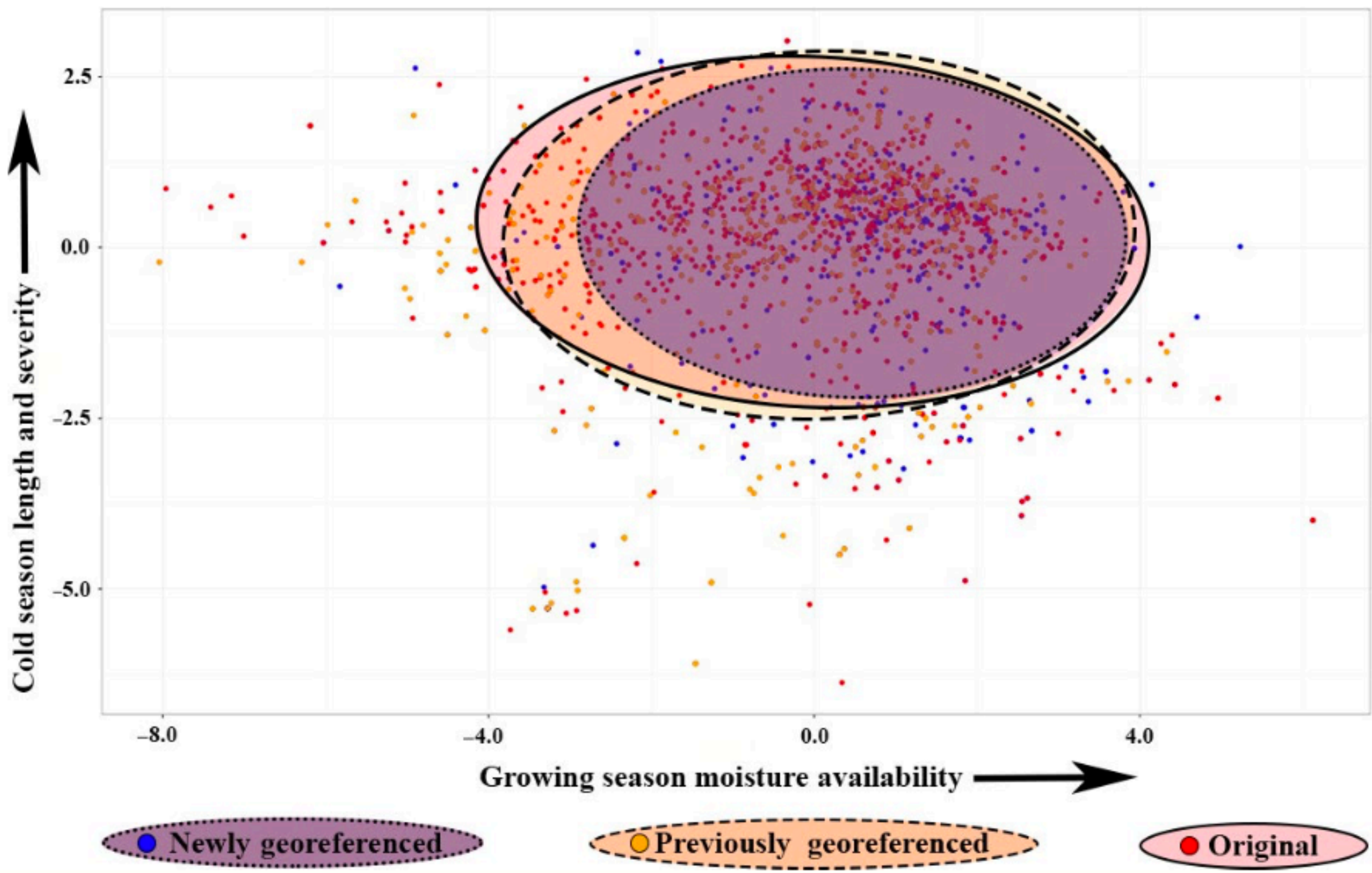
Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis

Bloom TDS, Flower A, DeChaine EG

Saxifraga austromontana

Endemic species of upper elevations in the Rocky Mountain Floristic Region.



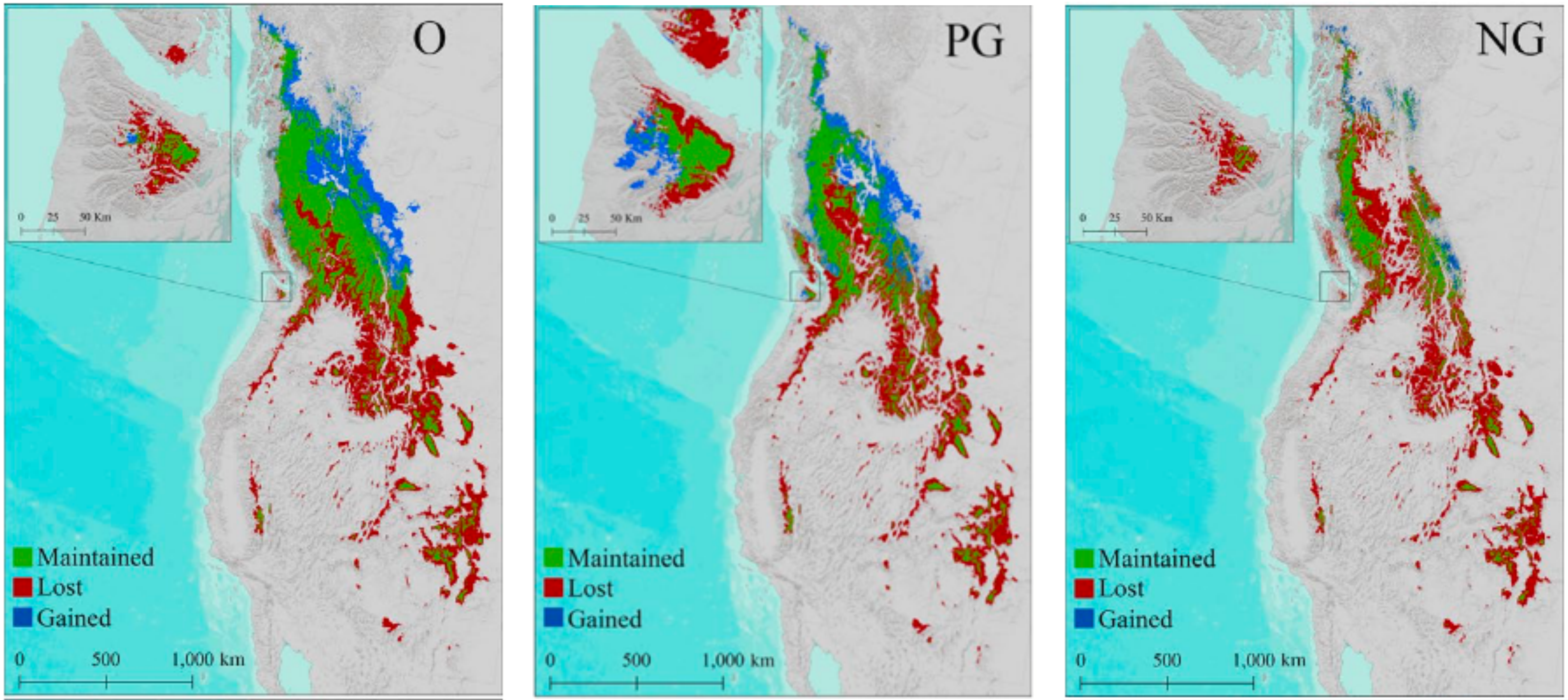

2.5
0.0
-2.5
-5.0
-8.0
-4.0
0.0
4.0
Cold season length and severity
Growing season moisture availability
Newly georeferenced
Previously georeferenced
Original

The authors developed 3 datasets:

O dataset, containing all available records of the species from several sources. It was edited to omit duplicate records and extreme outliers. Duplicate records across herbaria were found using accession numbers, GUID numbers, collector numbers, and site descriptions. Outliers were defined as occurrence records located very far outside of the known species range. Omission of outliers is common practice for building SDMs.

PG (Previously Georeferenced) dataset, including all records from the O dataset that explicitly state they have been georeferenced by other herbaria, using a variety of methods. Outliers and duplicates were omitted, as well as records with coordinate uncertainty >1 km.

NG (Newly Georeferenced) dataset includes all historical herbarium records from the O dataset which could be manually georeference to a 1-km, or finer resolution.



Since understanding present ad future distribution of a species is fundamental for conservation efforts, as well as for mitigation and adaptation to climate change, It is mandatory to develop models which are the most accurate as possible.
Thus, proper georeferencing practices should always be in place when preparing observation records for SDMs.

Data Cleaning

Data Quality

- data have the potential to be used in ways unforeseen when collected.
- the value of the data is directly related to the fitness for a variety of uses.
- "as data become more accessible many more uses become apparent." – Chapman 2005
- the GBIF Best Practices (Chapman and Wieczorek 2006) promote data quality and fitness for use.

Catching and Correcting

Geographic, and other errors

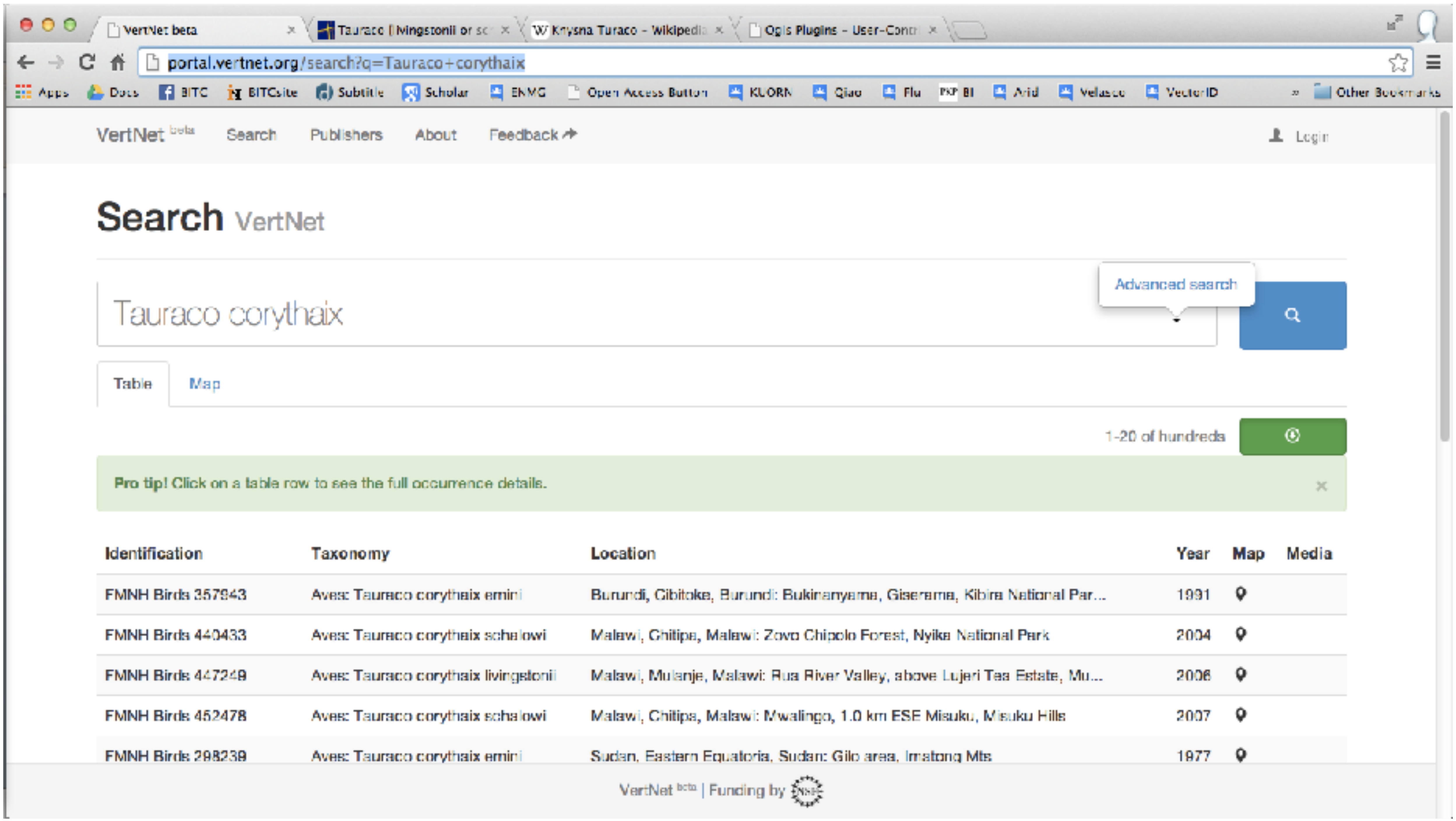
In

Biodiversity Data

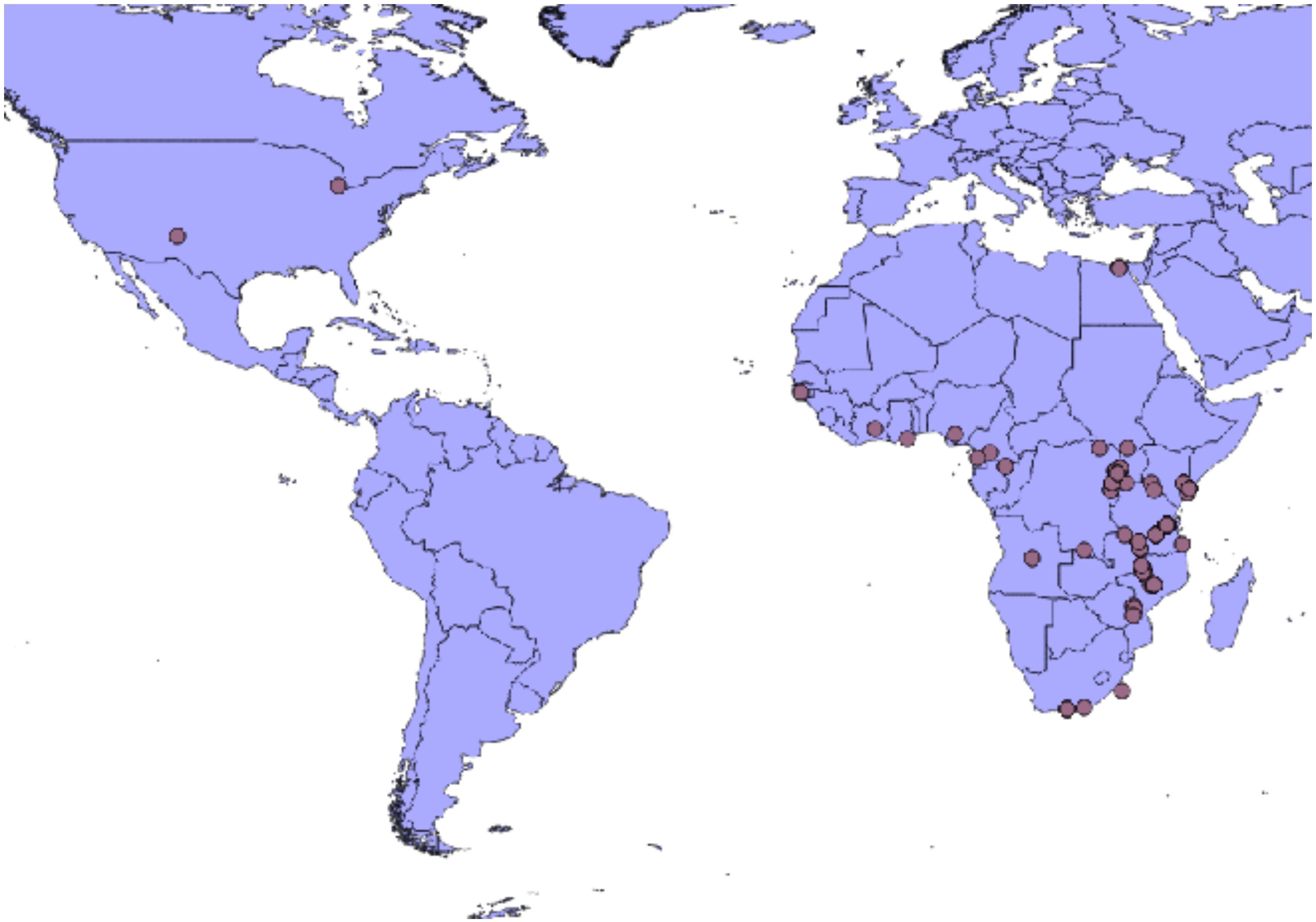
- Detect errors and potential errors
- Ideally, correct them in a way that is explicit and clear
- At least, flag them as dubious or potentially problematic, so that they can be treated as such
- Where possible, provide measures of confidence to qualify and enrich primary data

Tauraco corythaix Wagler, 1827

Data cleaning

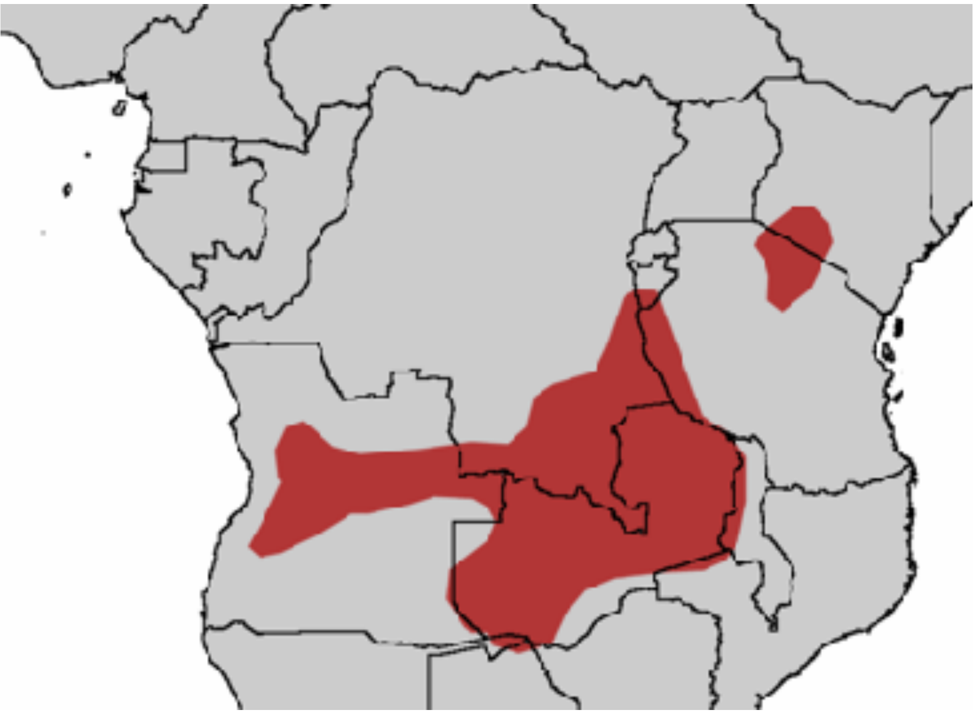


Data cleaning




Avibase - the world bird database
Bird checklists - taxonomy - distribution - maps - links
Home
My Avibase
Checklists
Search
Contribute
Translate
search avibase
Welcome guest | Login | Register |
Knysna Turaco (Tauraco corythaix) (Wagler, 1827)
summary taxon grid map eBird Wikipedia NatureServe ITIS Birdlife Flickr Audio More links
The Knysna Turaco (Tauraco corythaix), or, in South Africa, Knysna Lourie, is a large turaco, one of a group of African near-passerine birds. It is a resident breeder in the mature evergreen forests of southern and eastern South Africa, and Swaziland. It was formerly sometimes considered to be a subspecies of the Green Turaco of West Africa. The Livingstone's and Schalow's Turacos were once considered subspecies. Source: Wikipedia
© by fotoluluOnlineShop
photo ©fotolulu2012
(0 votes)
Photo powered by flickr.com.
Listen to recording
Stop recording
Order:
Musophagiformes
Family:
Musophagidae
English:
Knysna Turaco
Scientific:
Geographic range:
• Tauraco corythaix corythaix: Natal to w Zululand, s Swaziland and e Cape Province
• Tauraco corythaix phoebus: Humid forests of Transvaal and nw Swaziland
Source: Clements checklist 2012
Czech:
turako přilbový
German:
Helmturako
Danish:
Kapturako
Spanish:
Turaco de Knysna
Finnish:
etelänturako
French:
Touraco lourie
Italian:
Turaco verde del Sudafrica
Japanese:
eboshidori
Japanese:
エボシドリ
Dutch:
Knysnatoerako
Norwegian:
Natalturako
Taxonomic status:
Species status: full species
This taxon is considered a subspecies of Tauraco [corythaix or livingstonii] (sensu lato) by some authors
Your sightings
You must be logged in to view your sighting details. To register to myAvibase click here
Related taxa
Tauraco [livingstonii or schalowi]
• Tauraco schalowi
Tauraco schalowi (schalowi)
[Tauraco schalowi (chalcolophus)]
• Tauraco livingstonii
Tauraco livingstonii livingstonii
Tauraco livingstonii reichenowi
Tauraco livingstonii cabanisi
Tauraco [corythaix or livingstonii]
• Tauraco corythaix←
Tauraco corythaix corythaix
Tauraco corythaix phoebus
Other related concepts

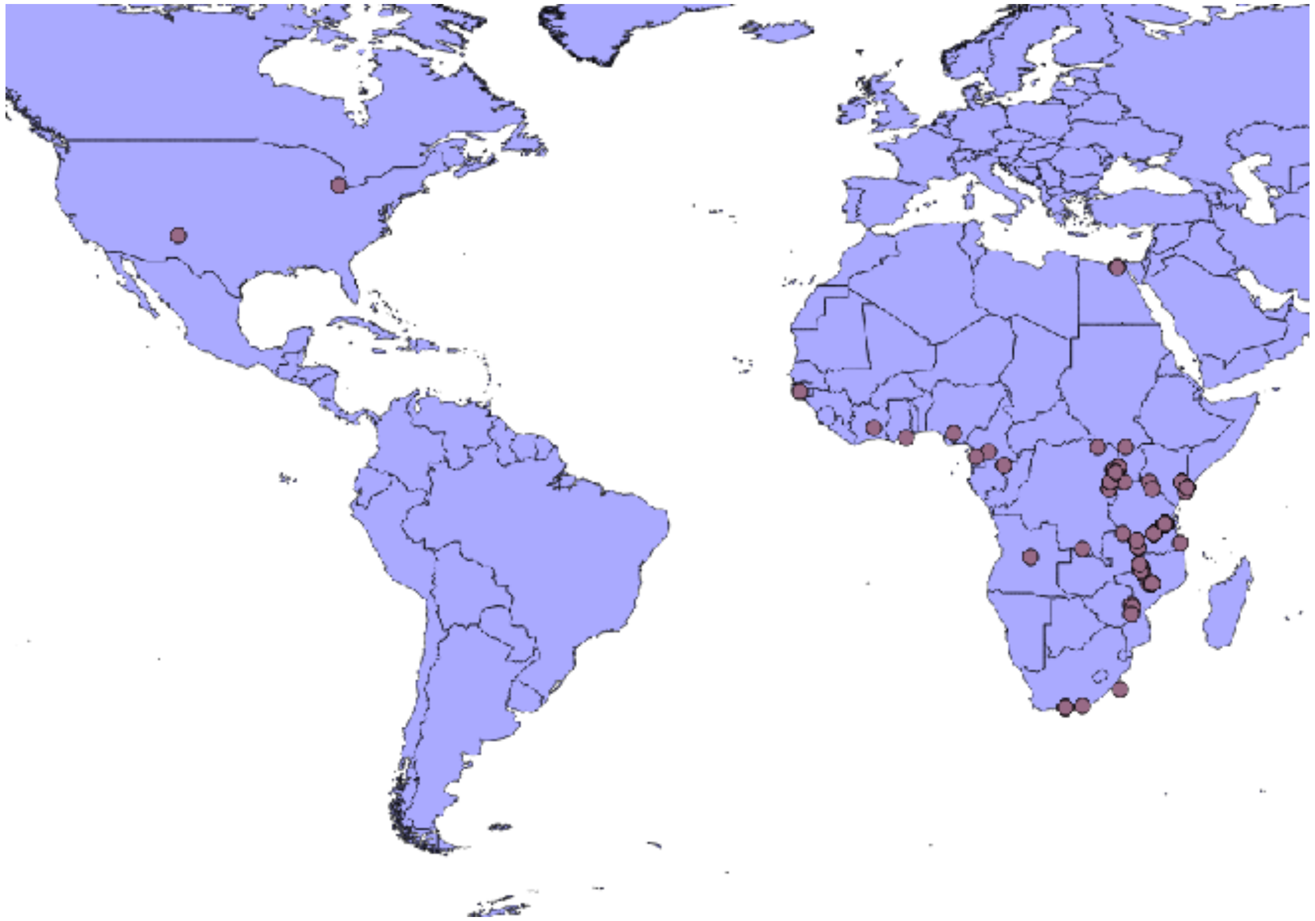
Data cleaning



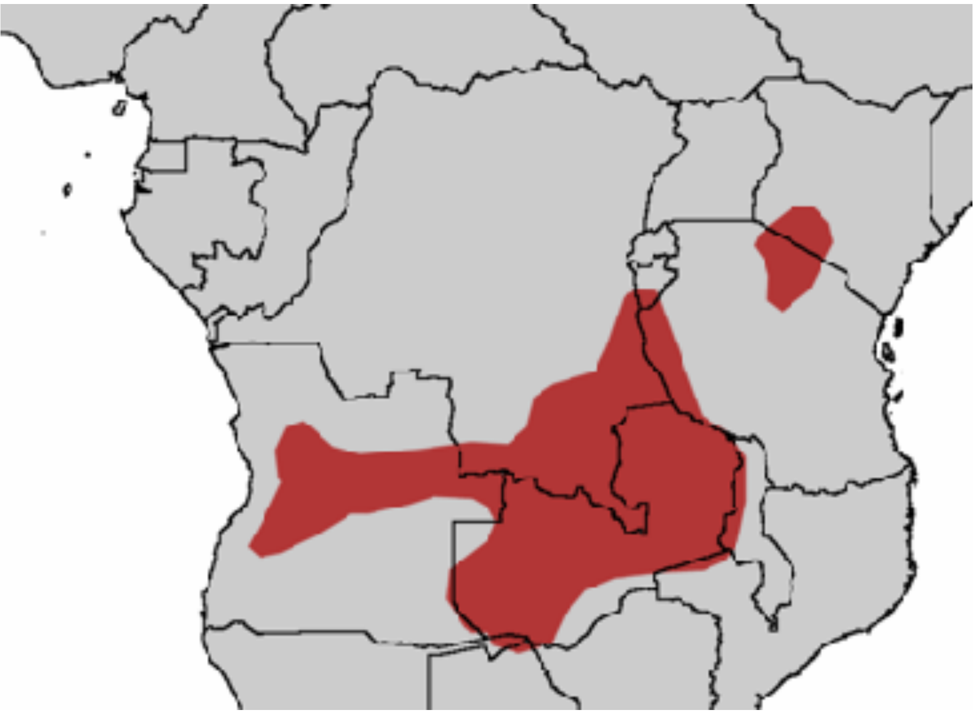
Tauraco schalowi

Tauraco livingstonii

Tauraco corythaix



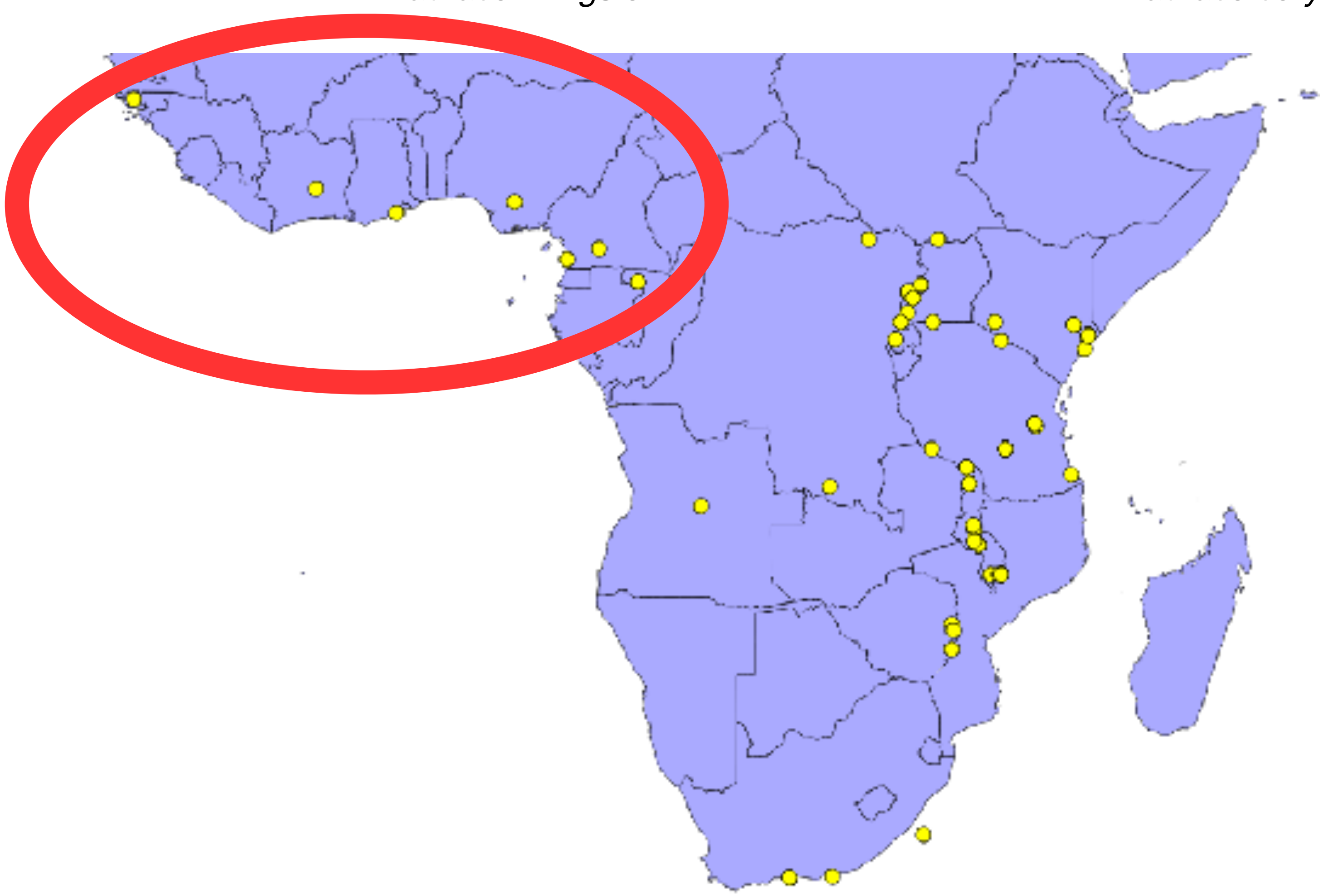
Data cleaning



Tauraco schalowi

Tauraco livingstonii

Tauraco corythaix



Error Detection, Data Cleaning, Error Flagging

- Biodiversity data are abundant, but can have LOTS of problems
  - Outdated taxonomy
  - Bad georeferences
  - Inconsistent ideas about “locality”
- Exploratory analyses and visualizations are crucial to any use
- Data cleaning must be part of any use of these data

Taxonomic checks

Data cleaning

Megaptera novaeangliae
Adult female, live

Off North Truro, MA, USA
42.101 N. 70.169 W

2010.09.29 21:47 GMT

Arturo H. Ariño
Aboard *Dolphin VI*
Canon Eos 450D, 200 mm lens

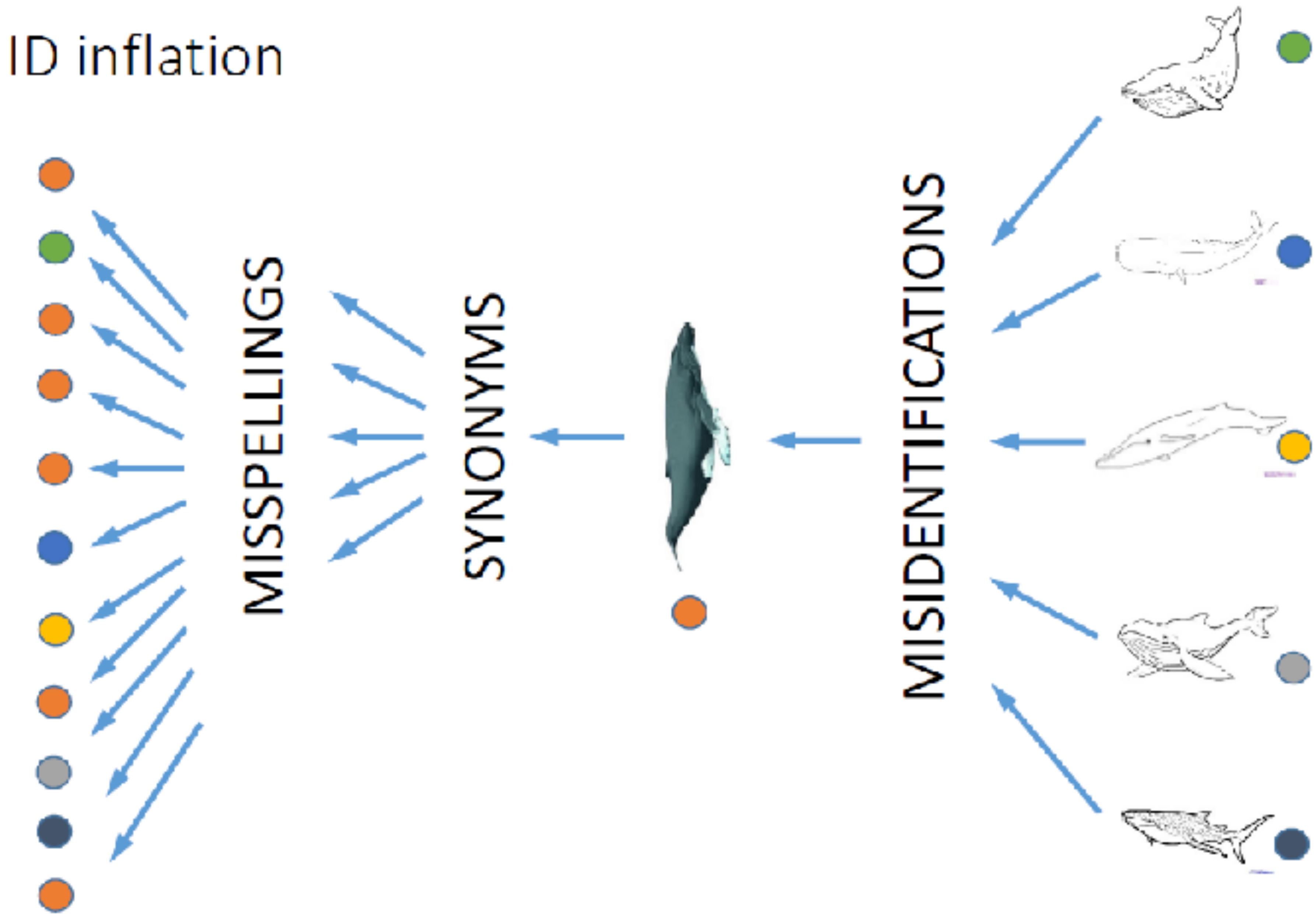
GBIF

How to also name a humpback whale

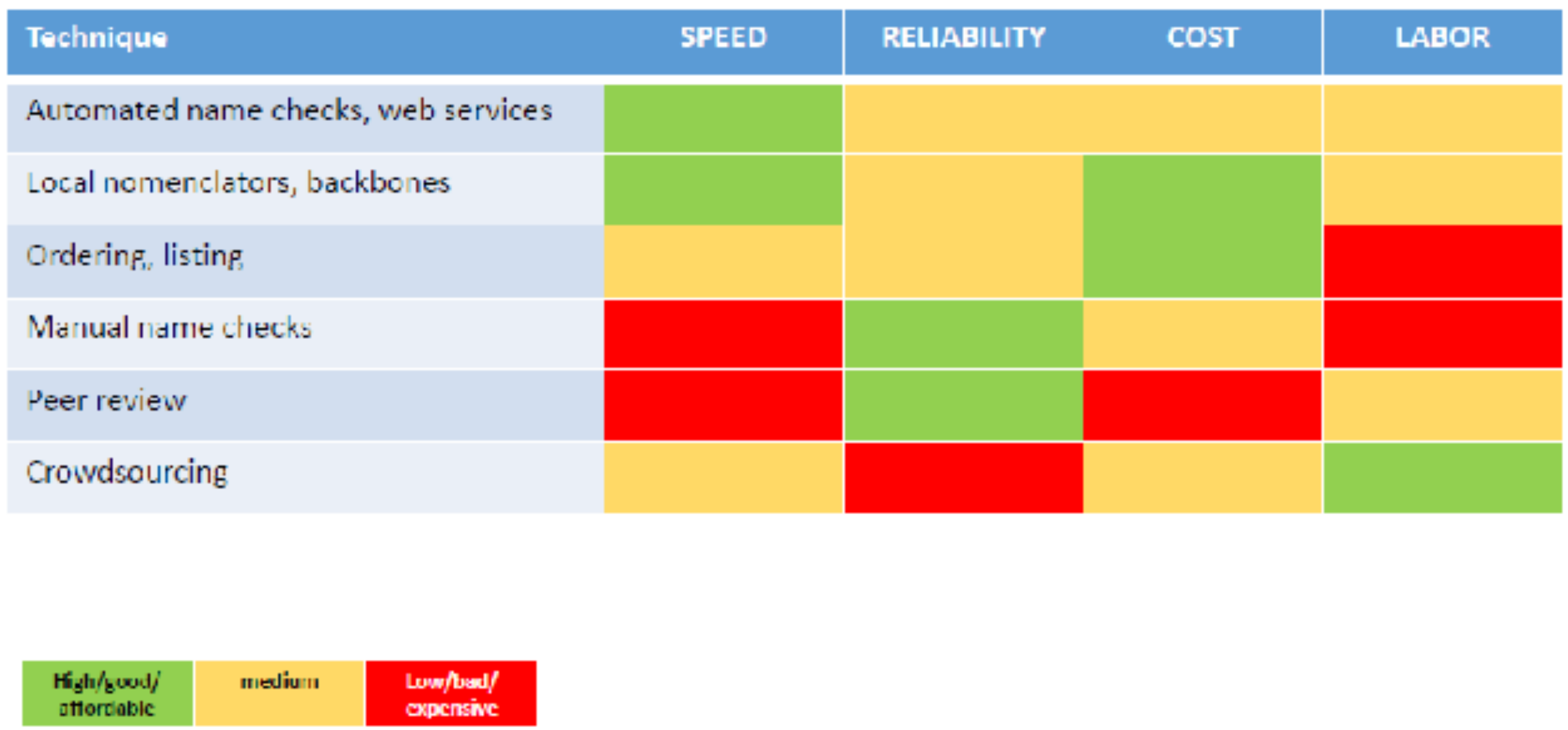
1. Balaena allamack Gray, 1846
2. Balaena atlanticus Hurdis, 1897
3. Balaena lalandii Fischer, 1829
4. Balaena longimana Rudolphi, 1832
5. Balaena nodosa Bonnaterre, 1789
6. Balaena novaeangliae Borowski, 1781
7. Balaena sulcata antarctica Schlegel, 1841
8. Balaenoptera astrolabe Pucheran, 1843
9. Balaenoptera capensis A. Smith, 1834
10. Balaenoptera leucopteron Lesson, 1842
11. Balaenoptera syncondylus A. Mueller, 1863
12. Kyphobalaena keporkak Van Beneden, 1868
13. Megaptera americana Gray, 1846
14. Megaptera antarctica Gray, 1846
15. Megaptera australis Iredale & Troughton, 1934
16. Megaptera bellicosa Cope, 1871
17. Megaptera boops Van Beneden & Gervais, 1880
18. Megaptera brasiliensis True, 1904
19. Megaptera braziliensis Cope, 1867
20. Megaptera burmeisteri Burmeister, 1866
21. Megaptera gigas Cope, 1865
22. Megaptera indica Gervais, 1883
23. Megaptera kusira Trouessart, 1904
24. Megaptera kuzira Gray, 1850
25. Megaptera lalandii (Fischer, 1829)
26. ... up to 46 synonyms

What can also go wrong?

MISIDENTIFICATIONS


ID inflation
MISSPELLINGS
SYNONYMS
MISIDENTIFICATIONS

Correction avenues



Correcting taxa: The Three Golden Rules

1. ALWAYS keep the original / verbatim data

2. ALWAYS record changes (as a new field)

3. ALWAYS follow rules 1 and 2.

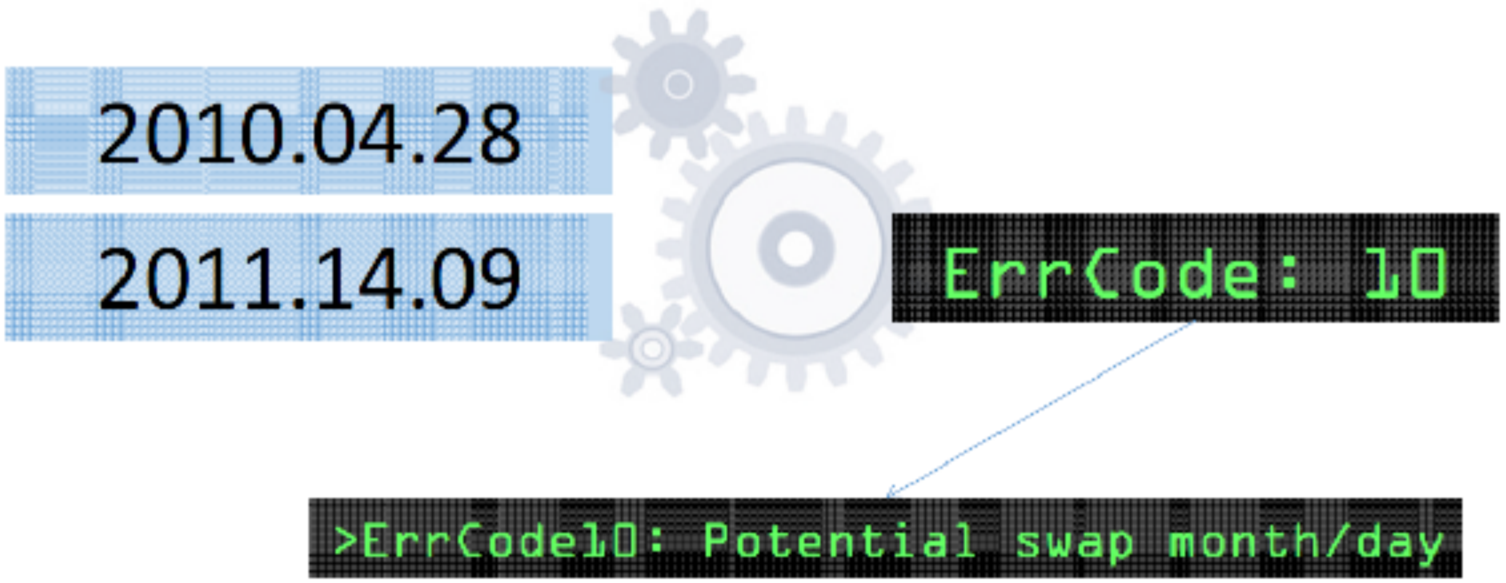
Now- What's in a date?

type	Order	HasYear	HasMonth	HasDay	HasHour	HasMinute	HasSecond	AMPM	TZ	UpTo	Range	LeadingZeroDay	leadingZeroMor	MonthAbbrev	DOW	Incidence
06/11/1904	XXY	Y	Y	Y	N	N	N	N	N	N	N	Y	Y	-	N	39,1%
10 Apr 1974	DMY	Y	Y	Y	N	N	N	N	N	N	N	Y	-	Y	N	21,6%
08/09/2003 02:11 PM	XXY	Y	Y	Y	Y	Y	N	Y	N	N	N	Y	Y	-	N	14,0%
1 Apr 1970	DMY	Y	Y	Y	N	N	N	N	N	N	N	N	-	Y	N	5,8%
-- Apr 1859	XMY	Y	Y	N	N	N	N	N	N	N	N	N	-	Y	N	3,5%
-- -- ----	XXY	N	N	N	N	N	N	N	N	N	N	N	-	-	N	3,3%
Locality: BLB	N/A	N	N	N	N	N	N	N	N	N	N	N	-	-	N	2,3%
11 September 2003	DMY	Y	Y	Y	N	N	N	N	N	N	N	N	-	N	N	2,2%
1608	Y	Y	N	N	N	N	N	N	N	N	N	N	-	-	N	1,8%
9 Jan 2006	DMY	Y	Y	Y	N	N	N	N	N	N	N	N	-	Y	N	1,6%
1909/03/21/1909/03/21	YMD	Y	Y	Y	N	N	N	N	N	N	Y	?	Y	-	N	1,1%
February 09, 2014 03:22	MDY	Y	Y	Y	Y	Y	N	N	N	N	N	N	-	N	N	0,7%
[date unknown]	N/A	N	N	N	N	N	N	N	N	N	N	N	-	-	N	0,6%
Aug 1925	MY	Y	Y	N	N	N	N	N	N	N	N	N	-	Y	N	0,2%
-- -- 192-	XXY	N	N	N	N	N	N	N	N	N	N	N	-	-	N	0,2%
April 28, 2013 12:25:11 PM PDT	MDY	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	-	N	N	0,2%
Mon Jun 18 2012 14:25:22 GMT-0400 (EDT)	MDY	Y	Y	Y	Y	Y	Y	N	Y	N	N	N	-	Y	Y	0,2%
-1984	Y	Y	N	N	N	N	N	N	N	Y	N	N	-	-	N	0,2%
2013-08-15 1906	YMD	Y	Y	Y	Y	Y	Y	Y	N	N	N	?	Y	-	N	0,2%
-- --- ----	DMY	N	N	N	N	N	N	N	N	N	N	N	-	-	N	0,1%
prior to 11 Jul 2007	DMY	Y	Y	Y	N	N	N	N	N	Y	Y	N	-	Y	N	0,1%
October 12, 2008	MDY	Y	Y	Y	N	N	N	N	N	N	N	N	-	N	N	0,1%
October 1953	MDY	Y	Y	N	N	N	N	N	N	N	N	N	-	N	N	0,1%
- Jul 1984	XMY	Y	Y	N	N	N	N	N	N	Y	N	N	-	Y	N	0,1%
99 XXX 9999	XXX	Y	N	N	N	N	N	N	N	N	N	Y	-	-	N	0,1%
2013-08-13 11:36am	YMD	Y	Y	Y	Y	Y	N	Y	N	N	N	?	Y	-	N	0,1%
2014-05-21 6:25:01 PM PDT	YMD	Y	Y	Y	N	N	N	Y	Y	N	N	?	Y	-	N	0,1%
2014-05-31 10:26 am PST	YMD	Y	Y	Y	Y	Y	N	Y	Y	N	N	?	Y	-	N	0,1%
2014-08-20 10:33AM	YMD	Y	Y	Y	Y	Y	N	Y	N	N	N	?	Y	-	N	0,1%
1884 1 1	YXX	Y	Y	Y	N	N	N	N	N	N	N	N	N	-	N	0,1%

2010.04.28

2011.14.09

YYYY.MM.DD



Data Cleaning
And
Error Detection

Generalities

- Error is universal in all data sets
- Data cleaning is a positive step that improves utility and fitness for use of a data set
- Existence of error SHOULD NOT be an excuse to do nothing or not to share data
- Error can be minimized for a particular application, and the consequences of its existence minimized
- No data set is clean… if someone tells you it is, don't believe them!

General Strategies

- Consistency is key
- Where multiple information sources are available within a data record, internal consistency is useful
- Where external information sources are available, external consistency can be even more powerful
- An iterative process of data exploration, visualization, and documentation of efforts

Flagging and Fixing Errors

- Very feasible to detect records that have high probability of holding errors
- Can be more difficult to fix them
- For one-off applications (e.g., an individual research project), can just eliminate them
- For archival applications (e.g., capturing and improving data records), must document, flag, correct, and document what was done
- BUT they may also be important records that teach real lessons about distributions and ecology of species

Environmental predictors

Environmental predictors are all those environmental variables which can influence the fitness of an organism in a given place.
They can be:

- Climate data
- Land cover and land-use data
- Geographic borders and the vector data
- Digital elevation data
- Geological data

There are a large number of available datasets ranging from the global to the local scale. They can be in the form of raster or vector layers, and can have continuous, or discrete values.

Continuous, and discrete variables

Environmental variables can be continuous or discrete.
Continuous variables are climate, and altitude, which attributes change in a continuous way - while more or less steeply - in the geographic space. Temperature can be an example. It changes continuously along any geographic gradient, latitudinal, longitudinal, or altitudinal.
Discrete variables are those which attribute change abruptly from one point to the other in the geographic space. An example can be geology. Especially in rocky areas, rocks can be of different origin, and thus have different properties. An extreme example can be a calcareous outcrop in the middle of a siliceous plateau. There is no transition between the two types of rocks.
In general, continuous variables are well represented by raster layers, while discrete varialbles are well represented by vector layers.

Continuous, and discrete variables