ISSN 1948-6596

data

Improving the use of information from museum specimens: Using Google Earth© to georeference Guiana Shield specimens in the US National Herbarium

Eduardo Garcia-Milagros and Vicki A. Funk

US National Herbarium, Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington DC, 20013-7012 USA
e-mail: eduadogarmi@gmail.com; http://botany.si.edu/bdg/index.html

Abstract. Data found on labels of museum collections have been useful in a variety of biodiversity studies. However, the georeferenced data available are often hampered by poor interpretation of label information and as a result are not as accurate, and therefore useful, as they might be. We have used Google Earth© as a geographic information system to improve the georeferencing of the data. Its user interface allowed us to make use of all the label information and to represent the coordinates more accurately, thus producing a better quality and more reliable dataset to be used in our studies. The quality, defined as "fitness for use", of the species-occurrence data generated, which is mostly affected by the values of accuracy and uncertainty associated to the coordinates, shows that uncertainly can be reduced. This method also allows us to show the power of examining georeferenced data from the stand point of 'all collections from an expedition" rather than "all collections from a single area." Type specimens housed at U.S. National Herbarium from the Guiana Shield were used in this work.

Keywords. museum collections, georeferencing, data quality, Google Earth, type specimen

Introduction

The specimen collections housed in museums and herbaria are a permanent record of a species at a given location on a specific date. The locality of a collection is stored as text on a specimen label. Georeferencing is the process of converting these locality descriptions into latitude/longitude coordinates which can be easily analyzed with GIS applications. These species-occurrence data, together with environmental variables are often used in various modeling methods, i.e., to plot existing data and predict the geographic distribution of species (e.g., Elith et al. 2006). These predictive distribution models are becoming an important tool in analytical biology, with applications in conservation and reserve planning, ecology, evolution, epidemiology, invasive-species management and other fields (Phillips et al. 2005); however, they depend on accurate coordinates.

Studies show that the data stored in the collections are often geographically, temporally, and taxonomically biased (Funk et al. 1999, ter Steege et al. 2000, Funk and Richardson 2002, Reddy and Davalos 2003). Although these studies suggest that collecting more data is necessary, the information behind these collections is of a high value. Gathering such data in databases and georeferencing them is a time-consuming and underappreciated task. However, once avaliable they are used, for instance, in establishing priorities for future expeditionary research and thus filling gaps in the data (Funk et al. 2005) and in regional conservation planning (Ferrier 2002, Chefaoui et al. 2005).

Recently there has been an increase in the availability of collections data (GBIF, TROPICOS, etc.) and an important consideration is how 'good' or reliable they are. Estimates of quality have been defined as "*fitness for use*" (Chrisman 1983) or "*fitness for potential use*" (English 1999) and Chapman (2005) describes how many factors may affect the quality of the data. In terms of geographic position of location, precision and accuracy are of concern and geographic data always

have an uncertainty value associated with them.

One goal of our research is to understand plant distributions across the Guiana Shield (Biological Diversity of the Guiana Shield Program, BDG). In order to do this we embarked on this project to apply the "*principles of the best georeferencing practices*" (Chapman and Wieczorek 2006) and investigate the use of Google Earth© as a GIS application for georeferencing and determine if its features could help improve the quality of the data. The full scope of the project involves checking the locations and uploading the data from all of the collections at US National Herbarium (US) beginning with those made by the BDG (see progress at http://botany.si.edu/bdg/expeditions.html). However, the sample data used here are from the type collection of the US which are important but provide the biggest challenge because of the lack of information.

Georeferencing type specimens from the Guiana Shield

All known species on earth have an official name. Typically that name consists of a genus, a specific epithet, and the name of the person(s) who described it. Usually each name is tied to a specimen that is housed in a recognized collection. These specimens are called 'types'. All type specimens from the Guiana Shield that are housed at US (ca. 3400 specimens) were used in this work.

When the US type specimen database was downloaded, it became clear that locality information varied from just country information on the old collections to precise GPS latitude/longitude coordinates on the most recent. Over time some older records had coordinates added. An examination of these data showed that during the process of entering them into the database, accidental errors had occurred. For instance, mistakes in typing a locality name made that location inaccessible in Gazetteers and changing label formats resulted in the loss of information. To avoid these and other pitfalls, we studied the types individually and used all the information that we had at the time to georeference them. Access to the original label through US Type Specimen Register Imaging Project (http://botany.si.edu/types/) has made this task easier.

The list below includes the fields that we found useful in the georeferencing process:

- *Locality names:* Towns, mountains, …
- *Elevation*.
- *Distances and heading:* 1 km N from…,
- *Habitats:* savanna, forest edge …
- *Expedition information:* base camp, intermediate camp …
- *Coordinates.*

We used Gazetteers for Guyana, Suriname, French Guiana and Venezuela (Defense Mapping Agency 1993a, 1993b, 1993c, 1993d) as a first approximation of the localities or when available, coordinates on the label. The set of coordinates (transformed to decimal degrees) were uploaded to Google Earth using EarthPlot, (free software: http://www.earthplotsoftware.com/) which allows easy plotting of large sets on Google Earth. In addition, we used maps of the Shield area, some published by different agencies and available through the BDG program map collection, and others from publications (i.e., Maguire 1945, 1948, Maguire and Reynolds 1955, Maguire and Wurdack 1959, Maguire 1981, Cowan 1952, Gleason 1931, Hitchcock et al. 1947, Huber 1995, Tate and Hitchcock 1930).

Below are two examples of how we have gathered and used the information to enhance the traditional georeferencing of collecting localities.

Example 1. (Figure 1) Some Types had coordinates on their labels but often these seemed to be an approximation. For example, the Type specimen of *Rhamnus marahuacensis* Steyermark and Maguire (Rhamnaceae) collected by Steyermark 126049 had coordinates that in Google Earth fell at the SE base of Cerro Marahuaca (Figure 1B). However, the label says that the specimen was collected at the 'summit of Cerro Marahuaca, in the Fhuif section at 2450-2500 m'. That area was found using a combination of the Google Earth location of the summit and the elevation; the coordinates were changed to reflect the more accurate location, resulting in a 15 km of distance from the original coordinates (Figure 1A).

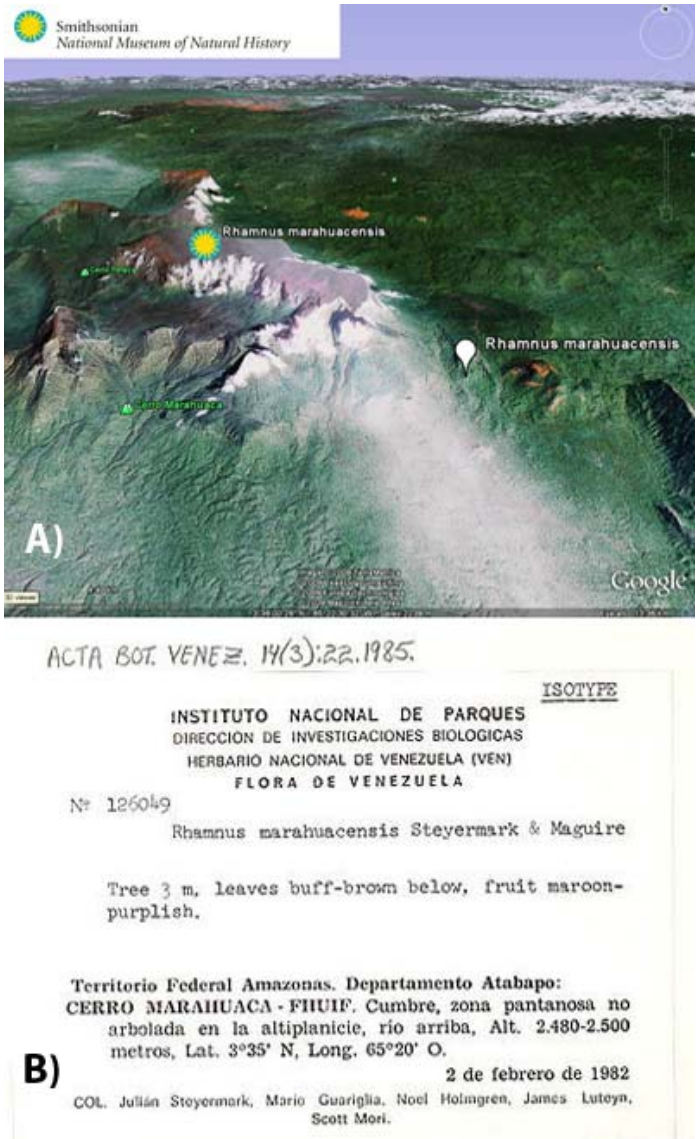

Figure 1. (A) Position of the coordinates (white placemark) and position after processing the label information. (B) Label of *Rhamnus marahuacensis* showing coordinates.

Example 2. (Figure 2) Some collections have references to such features as "*base camp, intermediate camp,…*" Sometimes these expeditions were the first exploration of an area, and they produced a large number of types. This happened with expeditions conducted in the amazing table top mountains or *tepuis* found on the Guiana Shield. An example is the expedition conducted by Maguire to Tafelberg, Suriname, in 1944. Tafelberg is an isolated sandstone Table Mountain representing a remnant of a *tepui*. All of the types from this area had been georeferenced in the database with the same coordinates, the summit. However, Maguire had published a map with the routes and the names that he gave to some of his collection localities (Figure 2B; Maguire 1948). Google Earth allows one to overlay images, such as the Maguire map, on its surface and therefore to create a georegistered version (Raes et al. 2009) of the map (Figure 2C). The type specimen of *Sagotia tafelbergii* Croizat (Euphorbiaceae; Figure 2E) collected by Maguire 24802 has "North Ridge" as a locality description, and having Maguire's map overlayed on the image from Google Earth allowed us to give more precise coordinates (Figure 2D) for the ca. 70 types housed at US that were collected from this expedition.

In addition to these examples, many other situations were encountered and locations subsequently corrected. To keep track of these changes new fields were added to the database. For instance, Example 1 had coordinates on its label so the new fields added to the database were: 1) initial source: coordinates label, 2) final source: Google Earth interface, 3) reason 1: label description, 4) reason 2: elevation.

One might ask how accurate the imagery and elevation data that Google Earth displays are. Google Earth uses WGS84 Datum as coordinate system and NASA Shuttle Radar Topography Mission data as Digital Elevation Model (although Google Earth may use different elevation data in some specific areas). We compared recent locality information recorded by two BDG collectors (D.H. Clarke's expedition to Mt. Ayanganna, Guyana, 2001 and K. M. Redden's expedition to Yatua River ,Venezuela, 2005) using GPS devices with Google Earth data; we found the average difference to be 50-100 m. The number of specimens and taxa studied and the coordinates provided are summarized in Table 1.

In total, the whole process of georeferencing the ca. 3400 Type specimens took eight months (appx. 100 specimens per week). The direct results of this study are available as place marks powered by Google Maps and Google Earth that can be downloaded and consulted on the website http://botany.si.edu/bdg/georeferencing.cfm.

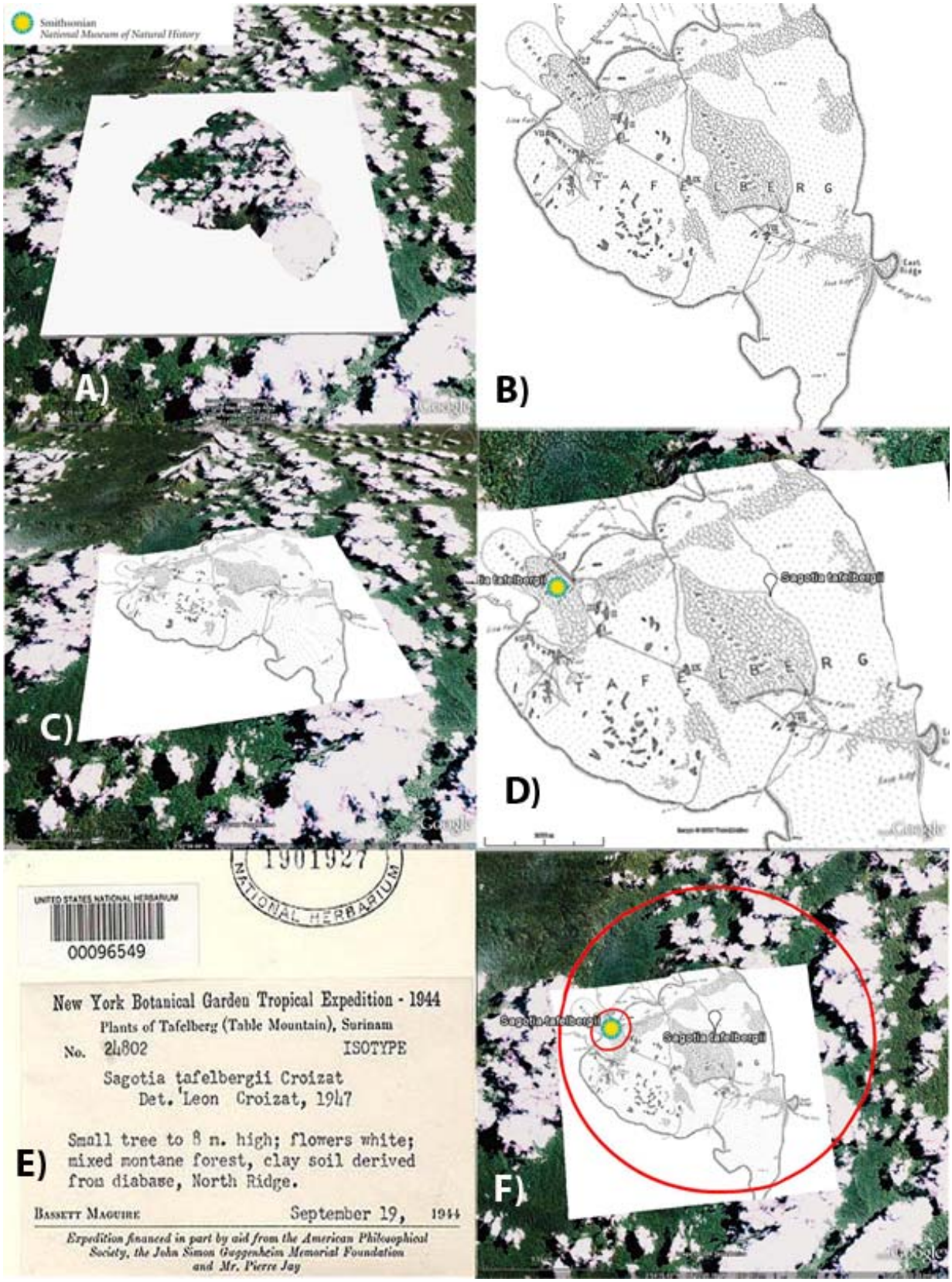

Figure 2. (A) Tafelberg cut section at 450 m asl on Google Earth. **(B)** Tafelberg expedition map (Maguire 1948) **(C)** Expedition map overlaid on GE. **(D)** Type georeferenced. **(E)** Label of *Sagotia tafelbergii*. **(F)** Point radius uncertainty as calculated according to Wieczorek (2004). The radius of the circle represents the maximum distance error for that locality, this example uses the type specimen of *Sagotia tafelbergii* (Example 2). The larger circle is the uncertainty if the Gazetteer coordinates for Tafelberg are used. The smaller circle represents the uncertainly when using the method described here.

The results are provided as a single coordinate pair assigned to each location. That does not mean that the collections georeferenced show the exact locality. All coordinates, even those that were obtained using a GPS device, have an uncertainty value associated with them. In fact, uncertainty is an inherent attribute of geographic information (Goodchild 2001). Using the protocol described above, we think that we have improved the quality of the data by increasing the accuracy and thus reducing the uncertainty. The uncertainty value is important because it can determine if the data are suitable for a particular analysis (Rocchini et al. 2011). For example, a plant locality description saying "*Banks of Potaro river*", might be useful for a research project on riparian vegetation, but not for a biodiversity survey of Kaieteur National Park or to predict distributions because, even though the Potaro River crosses the Park it is also outside of the Park and it crosses many vegetation types. Data from specimen labels have numerous sources of uncertainty: precision of the locality, unknown 'datum' information on maps, imprecise distance measurements or directional information, generalized or incorrect coordinates, etc. It can be challenging to calculate the uncertainty value when combining uncertainties from different sources. Chapman and Wieczorek (2006) provide different examples of calculating uncertainty depending on the locality information and proposed the point-radius method (Wieczorek et al. 2004) to represent uncertainty. This method describes each locality as a circle where the radius represents the maximum distance error for that locality, storing the uncertainty value as the length of the radius. The Biogeomancer Project (http://www.biogeomancer.org/) has an application for georeferencing localities providing the uncertainty values associated using this method. In the case of the Type of *Sagotia tafelbergii* (Example 2): if it is georeferenced using only the Gazetteer coordinates for Tafelberg its uncertainty is estimated as a circle with a 7.5 km radius which includes the whole mountain (Figure 2F). Overlaying Maguire's map (1945) on Google Earth allows the collection to be placed on the North Ridge thereby reducing the uncertainty to a circle with a 0.85 km radius (Figure 2F). In addition, Guo et al. (2008) recently proposed a probabilistic method to represent a locality with a polygon rather than a circle which tends to overestimate uncertainty.

For our studies, we do not include coordinates for those records with, in our viewpoint, high uncertainty values. In Table 1, the lower percentage ratio of the three Guianas is explained because the number of 'old' (historic) collections where inadequate locality information is common. For example, there are ca. 300 types collected from 1835-1844 by Robert and Richard Schomburgk with "*British Guiana*" or "*Banks of Essequibo*" as the only locality information. Despite

		Total Specimens	Total Taxa	Coordinates	No Coordinates	% Georefenced
French Guiana		235	217	130	87	59.9
Guyana		932	860	525	335	61.0
Suriname		258	239	156	83	65.3
Venezuela	***Amazonas***	1108	1015	986	29	97.1
	Bolivar	824	756	702	54	92.9
	Delta Amacuro	6	6	4	2	66.7
Total		3363	3093	2503	590	**80.9**

Table 1. Results of specimens georeferenced. The data showed on this table were taken at the moment of presenting the work (August 2010). The US National Herbarium database is being updated constantly and these numbers may vary at the present moment.

the extensive publication by van Dam (2002), it is still difficult to find a more precise locality and so they could not be mapped. Also excluded were collections with locality names that could not be found and others with inconsistencies in their information; these may be added in the future.

Concluding remarks

The advent of new GIS techniques and their potential for analyzing and interpreting the large quantity of data stored in specimen collections creates a challenge for researchers. Google Earth has proven useful in improving the quality of our data and we recommend its use to others for their georeferencing projects. It has an interface that easily allows overlying maps, drawing paths, adding information marks, measuring distances, checking the elevation of a point and moving a collection from one place to another. Such techniques provide a method to release previously incorrect or 'hidden' data often from ecosystems that are no longer extant.

There is still a lack of 'high resolution' imagery in Google Earth for many of the studied areas. Such updates would increase the applications of the tool in, for instance, overlaying predicted distribution polygons over real satellite imagery and modifying them according to identified habitats help in the production of more accurate vegetation maps, documentation of changes of habitat or land uses, or for planning future expeditions. Given the increased use of online databases it is critical that the data be checked and improved (see Hortal et al. 2007) so that it helps give accurate answer questions rather than unreliable results.

References

Chapman, A.D. (2005) Principles of Data Quality: version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Chapman, A.D. & Wieczorek, J. (eds). (2006) Guide to Best Practices for Georeferencing. Global Biodiversity Information Facility, Copenhagen.

Chefaoui, R.M., Hortal, J. and Lobo, J.M. (2005) Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. Biological Conservation, 122, 327-338.

Chrisman, N.R. (1983) The role of quality information in the long-term functioning of a geographic information system. Cartographica, 21, 79-87.

Cowan, R.S. (1952) Plant Explorations of G. Wilson-Browne, S. J. in British Guiana. I Kanuku Mountains. Brittonia, 7, 389–414.

van Dam, J.A.C. (2002) The Guyanan Plant Collections of Robert and Richard Schomburgk. Flora of the Guianas. Supplementary Series Fascicle 3. (ed. by M.J. Jansen-Jacobs). Royal Botanic Gardens, Kew, London.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129-151.

English, L.P. (1999) Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. New York, John Wiley & Sons, Inc.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Systematic Biology, 51, 331–363.

Funk, V.A., Zermoglio, M. F. & Nassir, N. (1999) Testing the use of specimen based collecting data and GIS in biodiversity exploration and conservation decision-making in Guyana. Biodiversity and Conservation, 8, 727–751.

Funk, V.A. & Richardson, K.S. (2002) Systematic data in biodiversity studies: Use it or lose it. Systematic Biology, 51, 303–316.

Funk, V.A., Richardson, K.S. & Ferrier, S. (2005) Survey-gap analysis in expeditionary research: where do we go from here? Biological Journal of the Linnean Society, 85, 549-567.

Defense Mapping Agency (1993a) Gazetteer of Guyana, May 1993. Second Edition; published by Defense Mapping Agency.

Defense Mapping Agency (1993b) Gazetteer of French Guiana, February 1993. Second Edition; published by Defense Mapping Agency.

Defense Mapping Agency (1993c) Gazetteer of Suriname, January 1993. Second Edition; published by Defense Mapping Agency.

Defense Mapping Agency (1993d) Gazetteer of Venezuela, April 1993. Second Edition; published by Defense Mapping Agency.

Gleason, H.A. (1931) Botanical results of the Tyler-Duida Expedition. Bulletin of the Torrey Botanical Club, 58, 277–344.

Goodchild, M.F. (2001) A geographer looks at spatial information theory. In Proceedings of COSIT 2001 (ed. by D.R. Montello). Lecture Notes in Computer Science (Berlin: Springer), 2205, 1–13.

Guo, Q., Liu, Y. & Wieczorek, J. (2008) Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. International Journal of Geographical Information Science, 22, 1067-1090.

Hitchcock, C.B., Phelps, W.H.Jr. & Galavis, F.A. (1947) The Orinoco-Ventuari Region, Venezuela. Geographical Review, 37, 525–566.

Hortal, J., Lobo, J.M. & Jiménez-Valverde, A. (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). Conservation Biology, 21, 853-863.

Huber, O. (1995) Geographical and physical features. In Flora of the Venezuelan Guayana: Volume 1. Introduction (ed. by J.A. Steyermark, P.E. Berry & B.K. Hoist), pp. 1–61. Timber Press, Portland, Oregon.

Maguire, B. (1945) Notes on the geology and geography of Tafelberg, Surinam. Geographical Review, 35, 563–579.

Maguire, B. (1948) Plant explorations in Guiana in 1944, Chiefly to the Tafelberg and the Kaieteur Plateau-I. Bulletin of the Torrey Botanical Club, 75, 56–115.

Maguire, B. & Reynolds, C. D. (1955) Cerro de la Neblina, Amazonas, Venezuela. Geographical Review, 45, 27–51.

Maguire, B. & Wurdack, J.J. (1959) Geographical record: The position of Cerro de la Neblina. Geographical Review, 49, 566–569.

Maguire, B. (1981) Introduction. In The Botany of the Guayana Highland- Part XI (ed. by Maguire, B. et al.) Memoirs of the New York Botanical Garden, 32, 1–3.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190, 231–259.

Raes, N., Mols, J.B., Willemse, L.P.M. & Smets, L.P.M. (2009) Georeferencing specimens by combining digitized maps with SRTM digital elevation data and satellite images: a Bornean case study. Blumea, 54, 162-165.

Reddy, S. & Davalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography, 30, 1719-1727.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G. & Chiarucci, A. (2011) Uncertainty in species distribution mapping and the need for maps of ignorance. Progress in Physical Geography, in press.

ter Steege, H., Jansen-Jacobs, M.J. & Datadin, V.K. (2000) Can botanical collections assist in a National Protected Area Strategy in Guyana? Biodiversity and Conservation, 9, 215–240.

Tate, G.H.H. & Hitchcock, C.B. (1930) The Cerro Duida Region of Venezuela. Geographical Review, 20, 31–52.

Wieczorek, J., Guo, Q. & Hijmans, R. (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science, 18, 745-767.

Edited by Joaquín Hortal