Integrating herbarium data into a geographic information system: requirements for spatial analysis

Ann F. Rhoads[1] & Laura Thompson[2]

Summary

Rhoads, A. F. & Thompson, L.: Integrating herbarium data into a geographic information system: requirements for spatial analysis. – *Taxon* 41: 43-49. 1992. – ISSN 0040-0262.

This paper addresses the importance of spatial resolution and data accuracy for the development and use of herbarium data with geographic information systems (GIS). Using the Pennsylvania Flora Database as an example, we discuss the variable quality of the spatial information in herbarium databases and present suggestions for standardizing spatial information for future data collection. GIS applications are discussed and examples of these applications are provided.

Introduction

Herbarium specimens are frequently the basis for systematic, floristic and biogeographic studies of plants. The many collections of specimens accumulated and maintained in herbaria throughout the world testify to their importance to the study of botany (Holmgren & al., 1990). With the recent availability of powerful yet inexpensive computers, the advantages of automating herbarium specimen records have become increasingly evident. Automated systems offer improved access to large quantities of data as well as expanded analytical potential.

In recent years a growing number of computerized, specimen-based botanical databases has been developed utilizing a variety of hardware and software environments (Gibbs-Russell & Arnold, 1989; Beaman & Regalado, 1989; Morin & al., 1989). Most botanical databases were developed with very specific uses in mind, including traditional herbarium management concerns such as providing inventories of specimens for quick reference, listing exsiccata and tracking loans.

As the number of databases increases so does the importance of developing standards to assure compatibility among systems and open the possibility of aggregating or comparing information from different sources. In response to this need, several efforts to standardize the format for taxonomic information have been developed (Anonymous, 1987b; Allkin & White, 1989). However, insufficient attention has been paid to the geographic, or spatial, component of herbarium databases. Although most data recording protocols contain some location information, usually the county or nearest named place, the inclusion of more specific collection locality information is highly variable.

As data are automated it is important to address the way location information is recorded in order to maximize the potential for spatial analysis of botanical data using

[1] Morris Arboretum of the University of Pennsylvania, 9414 Meadowbrook Ave., Philadelphia, PA 19118, U.S.A.

[2] Formerly of Expert Information Systems, School of Engineering and Architecture, Temple University, Philadelphia, PA 19122, U.S.A.

contemporary tools such as geographic information systems (GIS). In a GIS, data can be output as maps as well as in tabular form. Maps can reveal patterns or trends within the data which would be difficult to discern in tabular presentations (Burrough, 1986; Haslett, 1990). However, to realize the maximum potential of a GIS and to avoid misinterpretation, the data must have well-defined levels of spatial resolution and accuracy.

Spatial resolution and data accuracy

Resolution refers to the minimum size of a feature that can be mapped, e.g. one hectare or ten hectares. Accuracy reflects the indicated position of the mapped feature relative to its true location on the ground (see Burrough, 1986).

To be analysed spatially, botanical data must have locational information that can be related to a point on a map. The point might represent the location of a township (or other civil unit), a village (or other municipal unit) within the township, or the exact collection site recorded with geographic coördinates such as latitude and longitude. A botanical database which contains location information at each of these three levels of resolution could be used to generate many different maps to show herbarium information in a variety of ways. For example, data could be mapped by area

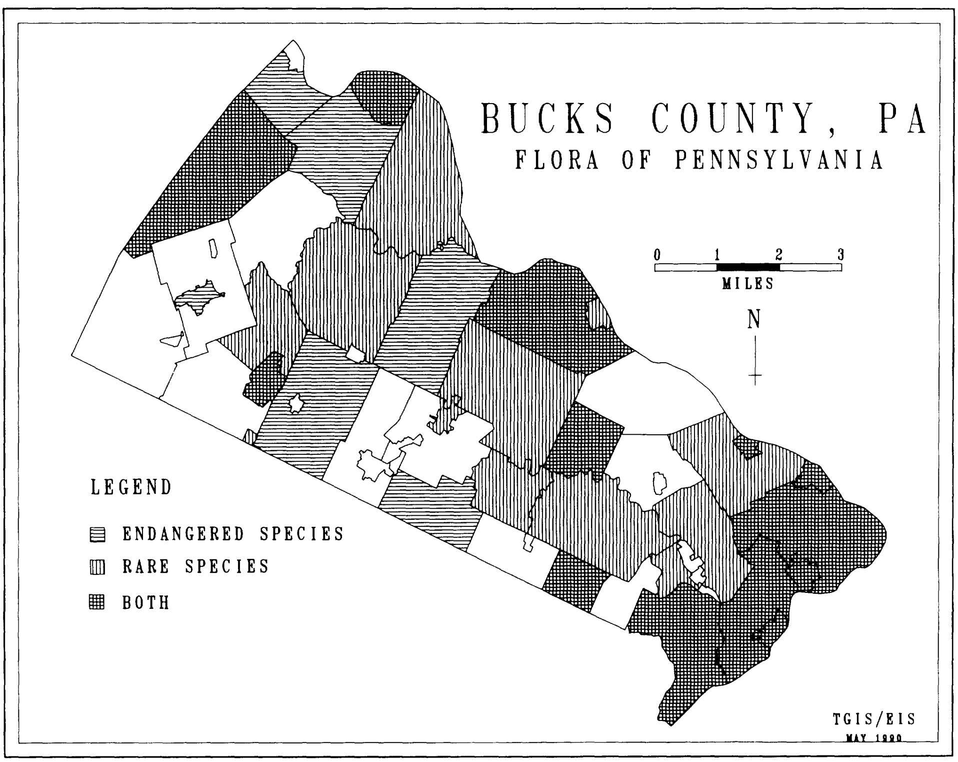

Fig. 1. Collection sites of endangered and rare plants of Bucks County, PA, displayed by township. Plant data from the Pennsylvania Flora Database developed and maintained at the Morris Arboretum of the University of Pennsylvania. Map prepared by Expert Information Systems, School of Engineering and Architecture, Temple University.

(township), by nearest place name (village), or by actual collection site. Mapping by collection site results in an accurate image, revealing the most detailed information about plant location. If the data are mapped by place name or area, the map presents equally accurate, but more generalized information. Fig. 1 and 2 illustrate this by showing the same data for a county in Pennsylvania displayed at the same level of accuracy but differing levels of resolution.

When location is recorded using precise measurements a variety of maps can be created according to how the data are to be used. Sensitive data can be masked by aggregating the point data into increasingly generalized levels of resolution. However, data gathered only at a low level of resolution (e.g. a county) can not be 'degeneralized' to show a more highly defined location (Burrough, 1986).

Using coordinates to describe a specimen collection site enables detailed spatial analysis of these data through correlation with other geographic data such as soil type, geologic formation or land use and land cover. Natural resource information is rarely mapped according to political units such as townships or counties. Rather, it is mapped according to a hierarchical classification system which incorporates many levels of resolution that can be aggregated to provide increasingly general information. For example, the Anderson Land Use Classification (Anderson & al., 1976) has three levels of spatial detail or resolution. The U.S.D.A. Soil Conservation Service

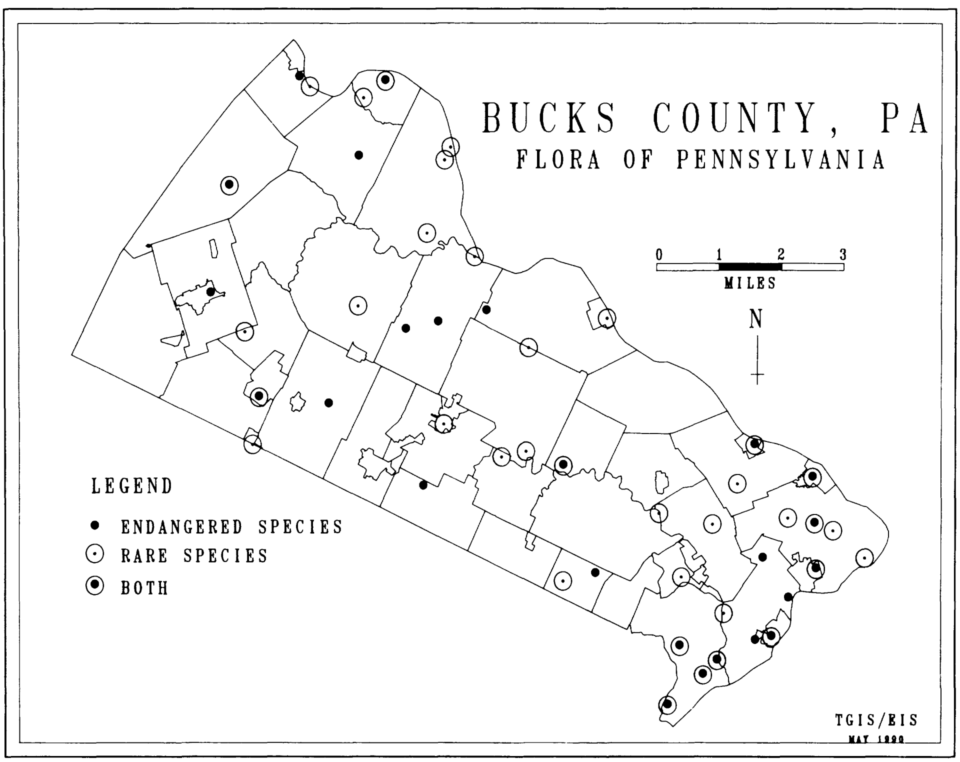

Fig. 2. Collection sites for endangered and rare plants of Bucks County, PA, displayed by specific location. Data source and map preparation as given for Fig. 1.

(Anonymous, 1975) maps soils by series which can be aggregated to form families, subgroups, great groups, suborders and orders, providing increasingly generalized information.

If botanical data were collected and recorded according to a similar multi-level, hierarchical spatial classification scheme there would be greater opportunity for spatial analysis and comparison with other data sets.

Traditionally herbarium specimen labels in Pennsylvania have included three, or sometimes four, levels of spatial information specifying the site at which the collection was made: state, county, nearest named place, and directions. Rarely are latitude and longitude included. In addition, especially in the case of older specimens (nineteenth and early twentieth century), the location information may be limited to county and a very general place name such as "Allegheny Mountains" which could apply to an area many hundreds of square kilometers in extent. Early in the development of the Pennsylvania Flora Database we decided to record all location information which was available and therefore devised an hierarchical set of data fields including: state, county, nearest named place, ancillary (directions), and latitude/longitude. All available information is recorded, and when information is not available the field is left blank.

Initially latitude and longitude values were obtained from a card file of approximately 8000 collection sites compiled by Edgar T. Wherry during the preparation of the *Atlas of the flora of Pennsylvania* (Wherry & al., 1979; Fogg, 1947). The Wherry list includes coordinates to the nearest half minute, information which he used to prepare the dot maps for the *Atlas*. While the accuracy of the information in the *Atlas* is judged quite high, the resolution of the data is limited by two factors. A minute is equal to approximately one mile in Pennsylvania, therefore coordinates rounded off to the nearest half minute refer to an area with a radius of one half mile. This degree of resolution was suitable for the *Atlas* where the dots as published are 3 miles in diameter (Fogg, 1947). However, for applications involving correlation with more detailed data such as soil series, it is not good enough.

Location to the nearest second could be more usefully compared. One source of latitude and longitude values which we have drawn on is the gazetteer of place names for Pennsylvania available in electronic form from the US Geological Survey (Anonymous, 1987a). Because this source includes latitude and longitude to the second, it should be more accurate than the Wherry values.

One disadvantage of this database, however, is the positional accuracy of the actual place names. Coordinate values for streams, for example, are determined from the mouth. This point may be a long way from the actual location where the specimen was collected. In addition, herbarium labels often do not contain accurate information regarding how far from the named place the specimen was actually collected. For these reasons caution is advised in relying solely on the gazetteer for determining latitude and longitude coordinate values for collection sites.

When the actual collection site can be located on a U.S. Geological Survey quad map, coordinates can be determined manually, although point accuracy is still affected by the scale (or resolution) of the map. In the future, improved accuracy and resolution may be obtained through the use of global positioning systems (GPS) to determine latitude and longitude values in the field at the time a specimen is collected.

Table 1. Suggested spatial hierarchy of resolution (adapted from Gibbs-Russell & Turner, unpublished).

Level I	Grid coordinates
Level II	Census District; Town Boundary; Management Area; District/Parish; Zip Code
Level III	Township; Section; Range
Level IV	County; Quadrangle; Physiographic Region
Level V	State/Province
Level VI	Country
Level VII	Major World Region

Classifying levels of resolution and accuracy

In mapping botanical data, care must be taken to avoid implying greater accuracy than the data can support. The presence of seemingly accurate coordinates in the database could be misleading if these values were not accurately obtained and do not in fact represent an actual collection site. To avoid misleading interpretations, a spatial hierarchy of resolution and accuracy could be developed. As Berry (1987) has pointed out "Those locations that are confidently derived should be differentiated from those predictions for locations less accurately defined . . . A spatial characterization of the accuracy of each variable [is necessary]."

We offer a spatial hierarchy in Table 1. The hierarchy has seven levels of resolution, from grid coordinates to major world region. Within each level of resolution varying levels of accuracy are possible. For example, grid coordinates obtained for a specimen mapped on a 1 : 24,000 topographic quad will be more accurate than the coordinate obtained for the same specimen mapped on a 1 : 50,000 topographic quad. Similarly, in most cases, a coordinate obtained for the same specimen using a GPS in the field will be more accurate still (see Slonecker & Hewitt, 1991, for further reading on GPS and other methods of measuring accuracy). In Table 2 we offer a set of codes to indicate the accuracy of the recorded location based on estimated proximity to the actual location on the ground. The actual code would consist of a combination of symbols from Tables 1 and 2. For example a code of I-2 indicates that specimen location was determined by grid coordinate, measured manually from a 1 : 24,000 topographic quad map, with an accuracy of ± 50 m. A code of II-4 would indicate a resolution of 'town' with the accuracy of location, measured from a road map, as ± 1 km.

A hierarchical classification scheme such as the one proposed in this paper will enable comparisons to be made between herbarium data and other geographical data.

Table 2. Suggested spatial accuracy codes.

Level of accuracy	method	code
± 10 m	GPS	1
± 50 m	1 : 24,000 (manual)	2
± 100 m	1 : 50,000 (manual)	3
± 1 km	road map	4

Since correlation becomes increasingly less accurate and more general as spatial resolution and data accuracy decrease, knowing the accuracy and resolution of the data will allow analysts to select records which are appropriate for a given application. It will also allow the use of data from different sources.

Conclusion

GIS technology has the potential to expand the applications of botanical databases through processes such as overlay of collection site data with natural resource and political data. The botanical GIS could become a powerful tool for analysing, verifying, improving, refining and adding information to botanical databases. It could allow spatial queries that a herbarium database alone can not, such as identifying new areas to search for rare species based on the occurrence of habitat characterizations matching those of known collection sites. In this way it could become an important tool in studies of endangered species and conservation efforts. At present there is a wealth of botanical data locked into databases which can not be easily analysed using GIS because of highly variable location information. Much of these data may never be suitable for detailed spatial analysis. However, by use of a coding system to indicate resolution and accuracy of individual records, appropriate uses can be determined. In the future, closer attention to standardizing the collection of location information for herbarium specimens will greatly enhance the usefulness of the resulting databases.

Acknowledgements

The work was funded by a joint grant from The Pew Charitable Trusts to the Morris Arboretum of the University of Pennsylvania for the Flora of Pennsylvania project and to the Missouri Botanical Garden for the Flora of North America project.

We thank Dr. John Radke, University of California, Berkeley; Expert Information Systems, School of Engineering and Architecture, Temple University; and Dr. Robert Wirtshafter and Robert Noland of The Center for Energy and the Environment, University of Pennsylvania for their assistance. Dr. Peter August, University of Rhode Island, kindly read and provided comments on a second draft of this paper.

Literature cited

Allkin, R. & White, R. J. 1989. *XDF: A language for the definition and exchange of biological data sets. A working manual, Version 3.2.* Taxonomic Databases Working Group (TDWG), [without place].

Anderson, R. J., Hardy, E. E., Roach, J. T. & Witmer, R. E. 1976. A land-use and land-cover classification system for use with remote sensor data. *U.S. Geol. Surv. Profess. Pap.* 964.

– 1975. A basic system of soil classification for making and interpreting soil surveys. *U.S.D.A. Handb.* 436.

– 1987a. *Geographic names information system.* U.S.D.I., United States Geological Survey, Reston, VA.

Anonymous, 1987b. *International transfer format for botanic garden plant records, plant taxonomic database standards. Version 01.00.* Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh, PA.

Beaman, J. H. & Regalado, J. C. 1989. Development and management of a microcomputer specimen-oriented database for the flora of Mount Kinabalu. *Taxon* 38: 27-42.

Berry, J. K. 1987. Computer-assisted map analysis: potential and pitfalls. *Photogrammetric Engin. Remote Sensing* 53: 1405-1410.

Burrough, P. A. 1986. *Principals of geographic information systems*, Oxford University Press, Oxford.

Fogg, J. M. 1947. Some methods applied to a state flora. *Contr. Gray Herb.* 165: 121-132.

Gibbs-Russell, G. E. & Arnold, T. H. 1989. Fifteen years with the computer: assessment of the "Precis" taxonomic system. *Taxon* 38: 178-195.

Haslett, J. R. 1990. Geographic information systems: a new approach to habitat definition and the study of distributions. *Trends Ecol. Evol.* 5: 214-218.

Holmgren, P. K., Holmgren, N. H. & Barnett, L. C. 1990. Index herbariorum. Part 1. The herbaria of the world, ed. 8. *Regnum Veg.* 120.

Morin, N. R., Whetstone, R. D., Wilken, D. & Tomlinson, K. L. 1989. *Floristics for the 21st century.* Missouri Botanical Garden, St Louis, MO.

Slonecker, E. T., Hewitt, M. J. 1991. Evaluating Locational Point Accuracy in a GIS Environment, *Geo Info Syst.*, June 1991.

Wherry, E. T., Fogg, J. M., & Wahl, H. A. 1979. *Atlas of the flora of Pennsylvania.* Morris Arboretum, University of Pennsylvania, Philadelphia, PA.