Combining Herbarium Data with Spatial Data: Potential Benefits, New Needs

M. E. Barkworth[1] and J. McGrew[2]

[1]Intermountain Herbarium, Department of Biology, Utah State University, Logan, Utah, U.S.A. 84322-5305; [2]Department of Landscape Architecture and Environmental Planning, Utah State University, Logan, Utah, U.S.A. 84322-4005, e-mail: mary@biology.usu.edu

Abstract: *Herbarium* specimens are, potentially, a rich source of information on the past and present distribution of species. For their potential value to be realized, information from the specimen labels must first, if feasible, be georereferenced, and then entered into a database. The Global Biodiversity Information Facility (GBIF) has devised protocols for making available information from all the world's herbaria (and other natural history collections). In this paper we demonstrate how such data can be combined with spatial data from other resources to determine the distribution and evaluate the ecological characteristics of different species. The demonstration makes evident that the data in GBIF are, at present, biased, more data being available from Europe and North America than from other parts of the world. To overcome this, every effort must be made to improve the human and financial resources available to herbaria in order to broaden the participation in GBIF.

Keywords: plant distributions; Global Biodiversity Information Facility, Herbarium databases, ecology

The world's 3.200 + registered herbaria (http://sciweb.nybg.org/science2/IndexHerbariorum.asp) are a rich and verifiable source of distributional information that can be used to evaluate the ecological characteristics of individual taxa. Many herbaria are now making information from their collections freely available through GBIF, the Global Biodiversity Information Facility (2005). The mission of GBIF is making information from the world's natural history collections, including herbaria and genebanks, freely accessible via the Internet. In this paper, we demonstrate how such information might be used.

GBIF has developed standards, structures, and tools for making herbarium data available. It is working with the genebank community on the development of appropriate standards and tools for sharing the data associated with genebank accessions. Until these are developed, information about plant biodiversity and distributions will be limited to data from herbarium specimens.

The task of entering specimen records into herbarium databases, has only begun. The cost of doing so is huge. It is even greater if retrospective georeferencing (determining the latitude and longitude of the collection sites) of specimens that were not georeferenced at the time of collection is considered the first step in databasing. Depending on the information provided on the label, this can easily add 5 minutes to the task of databasing a specimen record. Nevertheless, it is worth doing, for records with such information can be analyzed with respect to the vast array of spatially-related data becoming available, e.g. elevation, precipitation, soil salinity, and maximum and minimum temperatures. In this paper, we demonstrate one use of the information available via GBIF, that of comparing the ecological range of species. We also demonstrate some limitations of the data currently available from GBIF.

MATERIALS AND METHODS

We selected three species for this demonstration: *Elymus caninus*, *E. glaucus*, and *Pseudoroegneria spicata. Elymus caninus* is a widespread Eurasian species; *E. glaucus* and *P. spicata* are native to western North America. We downloaded georeferenced data from GBIF on May 21, 2005 and used it to plot the distribution of each species with ArcView 3.3 (ESRI 1998). Data for *E. caninus* were displayed using Robinson projection; those for *P. spicata* and *E. glaucus* used an Albers Equal Area projection.

The georeferenced data for *Pseudoroegneria spicata* and *Elymus glaucus* were used to determine the elevation, precipitation, and minimum and maximum temperature of the collection sites. These were then summarized in Excel (Microsoft 2003). All data were reprojected to Albers Equal Area projection for display.

RESULTS

We obtained 5769 georeferenced records for *Elymus caninus*, 240 for *Pseudoroegneria spicata,* and 275 for *E. glaucus*. Most records for *E. caninus* (5432) were provided by the British Botanical Society; most of those for *P. spicata* (236) and *E. glaucus* (186) came from the Intermountain Herbarium of Utah State University.

Elymus caninus

The map for *Elymus caninus* (Figure 1) reveals one of the limitations of the data currently available from GBIF. The species extends from Europe to western China (Tutin *et al.* 1980), but the georeferenced data in GBIF on May 27, 2005, show it as a primarily British and Scandinavian species, with scattered records from elsewhere in western Europe. This reflects the location of the three major providers: UK NBN DiGir Provider (5432 records), GBIF Sweden Provider (156 records), and Süddentsche Bergwalder (72 records). There were 965 records for *E. caninus* in GBIF that were not georeferenced. These provided a somewhat more accurate portrayal of the species' distribution, adding Pakistan and China to the countries for which there were records, for a total of 13 countries, in 10 of which the species is probably native. Nevertheless, it was clear that when this paper was prepared for presentation, GBIF's data providers were primarily western European and that the distributional information in the data reflected this bias.

The map shows a few anomalous records of *Elymus caninus*, three in the U.S. and two in Australia. The Australian specimens stated in the locality column that they were grown on experiment plots. Two of the U.S. specimens were also of cultivated specimens. The third may reflect use of a different taxonomic treatment. Hitchcock (1960) treated

Figure 1. Distribution of *Elymus caninus* based on georeferenced records in the Global Biodiversity Information Facility on May 27, 2005. The majority of the available records were provided by the UK NBN DiGIR provider (5432), the second largest number by the GBIF Sweden Provider (156)

Figure 2. Distribution of *Pseudoroegneria spicata* (triangles) and *Elymus glaucus* (dots), based on georeferenced records in the Global Biodiversity Information Facility on May 27, 2005. Most of the records came from Utah State University's Intermountain Herbarium

North American *Agropyron trachycaulum* (~*Elymus trachycaulus*) as a subspecies of *Agropyron caninum* (~*Elymus caninus*). Because he wrote the standard flora for the Pacific Northwest of North America, his treatment is still reflected on many herbarium specimens. Because GBIF shows the origin of each record, it is possible to send a message to the institution involved asking for an explanation of anomalous records. Herbaria generally welcome the opportunity such queries give them for correcting errors.

By July 15, 2005, the number of records in GBIF for *Elymus caninus* had increased to 9389, with 8761 being georeferenced. The number of countries represented had increased to 29, but the only georeferenced records from east of 23°E were of plants cultivated in Australia. Changing this bias in the data from GBIF's available records will require both technological and human investment in the infrastructure of the herbaria in the countries concerned. The investment involves more than the provision of computers and network access. Effective databasing requires assigning a unique identification number to each specimen and, possibly, modifying the work flow for preparation of new specimens – in addition to reviewing, possibly georeferencing, and then databasing existing specimens. Nevertheless, it is only through such investment that the promise of GBIF can be realized.

Pseudoroegneria spicata **and** *Elymus glaucus*

Pseudoroegneria spicata and *Elymus glaucus* are western North American species (Figure 2) with similar geographic distributions, the main difference being that *E. glaucus* extends to the coast whereas *P. spicata* is an inland species. Both also grow in British Columbia, Canada, but the avail-

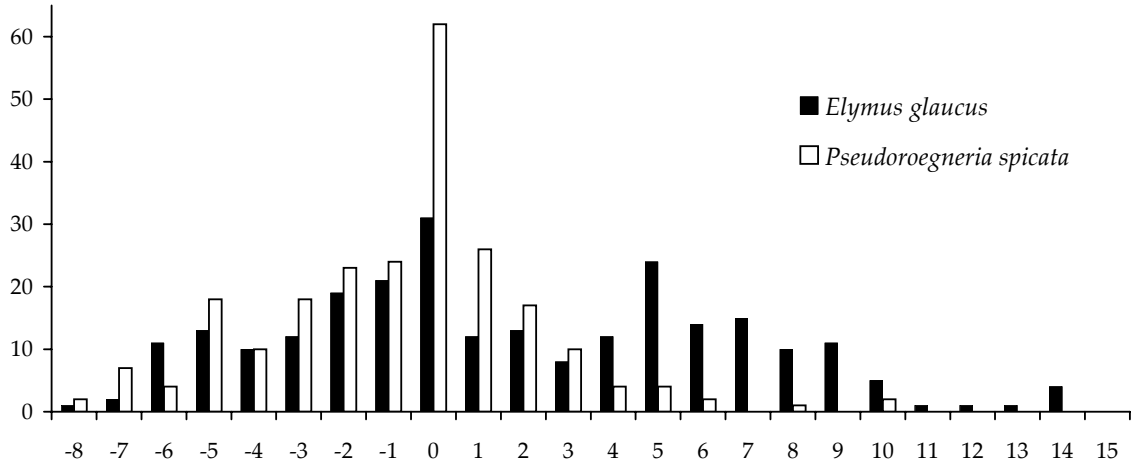

Figure 3. Comparison of the minimum temperatures at sites where *Elymus glaucus* and *Pseudoreogneria spicata* grow, based on georeferenced records in the Global Biodiversity Information Facility on May 27, 2005. For the source of records and temperature data, see text

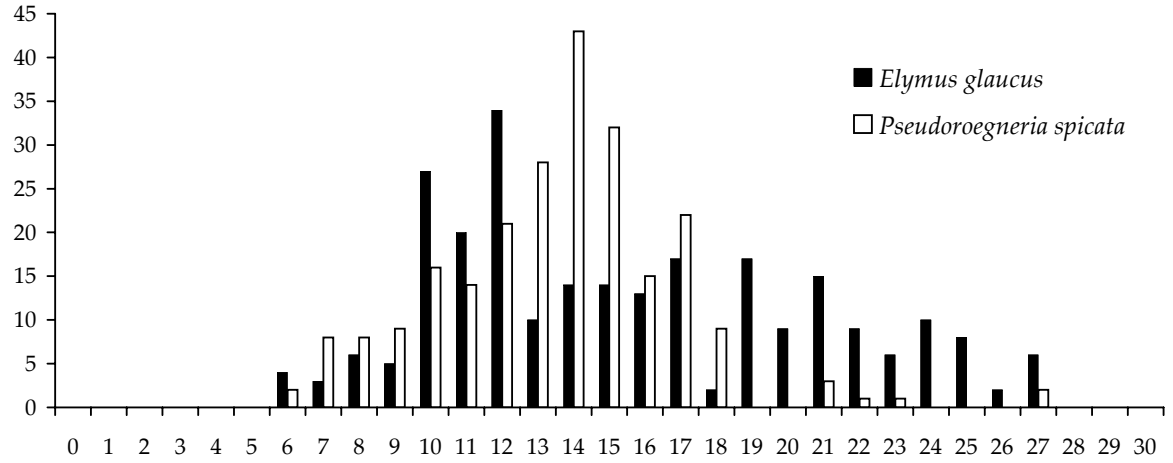

Figure 4. Comparison of the maximum temperatures at sites where *Elymus glaucus* and *Pseudoreogneria spicata* grow, based on georeferenced records in the Global Biodiversity Information Facility on May 27, 2005. For the source of records and temperature data, see text

ability of GIS data from different countries varies, as do the methods by which they were produced and the extent to which these methods are documented. These factors make it easier to work with spatial data from a single country. Consequently, we worked only with data from U.S. collection sites. As with *Elymus caninus*, there are many records of both taxa in GBIF that lack georeferenced data. They do not, however, expand the U.S. distribution shown, indicating that, for these two taxa, the georeferenced data provide a reasonably accurate portrayal of the distribution of the two species.

Examination of the ecological data (Figures 3–6) for *Pseudoroegneria spicata* and *Elymus glaucus* show some differences between the two species, despite the fact that the resolution of the spatial data is only 75 km. *Pseudoroegneria spicata* sites tend to be at higher elevations, and have higher maximum temperatures and lower precipitation than collection sites for *E. glaucus.* The somewhat bimodal distribution of the values for *E. glaucus* suggests that the taxonomic treatment and/or biology of the species would be worth investigating. It suggests that there may be two distinct ecotypes present. There is, of course, no guarantee that such ecotypes, should they exist, could be identified morphologically. The Global Biodiversity Information Facility also permits storing elevation data from herbarium records. Such data would be more reliable than data estimated from retrospective georeferencing

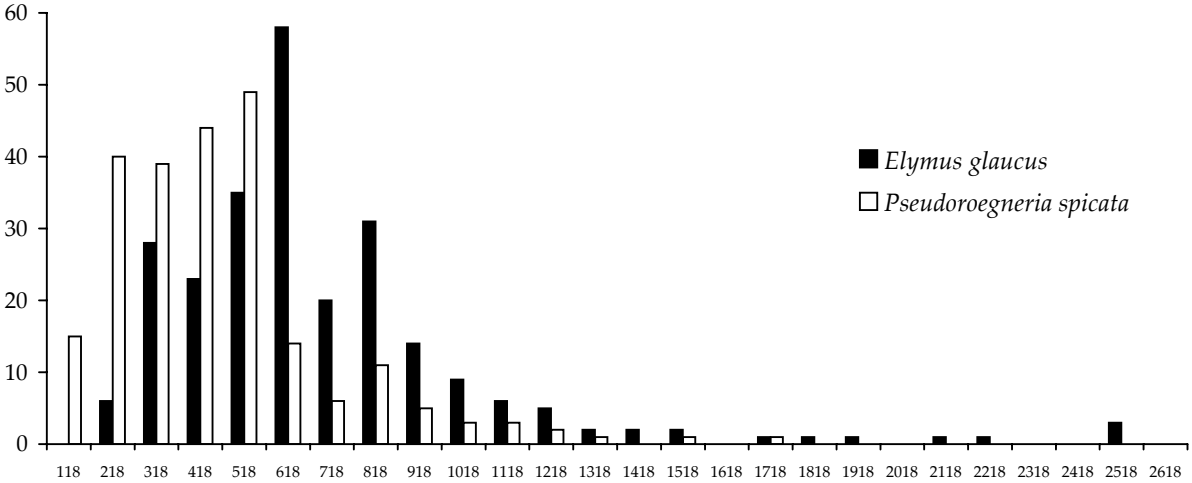

Figure 5. Comparison of the precipitation at sites where *Elymus glaucus* and *Pseudoreogneria spicata* grow, based on georeferenced records in the Global Biodiversity Information Facility on May 27, 2005. For the source of records and precipitation data, see text

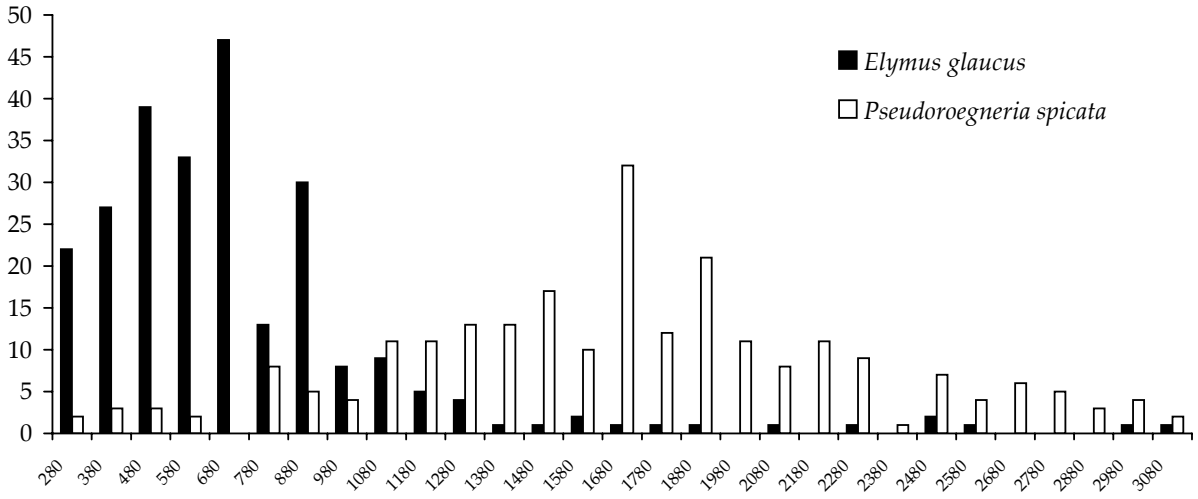

Figure 6. Comparison of the elevation at sites where *Elymus glaucus* and *Pseudoreogneria spicata* grow, based on georeferenced records in the Global Biodiversity Information Facility on May 27, 2005. For the source of records and elevation data, see text

and might even be used to improve the accuracy of such georeferencing. Somewhat to our surprise, none of the records for either *P. spicata* or *E. glaucus* included elevation data.

The range in ecological values for each species is substantial. This may reflect the lack of precision in the georeferencing. Most herbarium specimens, including those in the Intermountain Herbarium, do not have georeferenced data on the label. There are tools for estimating such data. In the western U.S., useful tools include the Geographic Names Information System website (http://www.gnis.gov/), StreetAtlas (DeLorme 2004), and the TRS2LL website (http://www.geocities.com/jeremiahobrien/trs2ll.html). The results are, of course, only estimates and it is difficult to provide a meaningful estimate of their inaccuracy. The reason for this, however, lies mostly in the impossibility of determining the accuracy of the label data. A label may read "Five miles west of Logan, Utah". This does not indicate where the five miles were measured from, nor how much accuracy can be associated with "five miles" and "west". In mountainous areas, such as western North America, inaccuracies in georeferencing can lead to significant differences in site characteristics. With the development of (relatively) inexpensive GPS units, more and more specimens are coming in with accurate GPS data, but the value of herbaria data resides, at least in part, in their ability to show us the past distribution of species. Whether the inaccuracy of the locality data associated with older specimens is acceptable must depend on the use being made of the data.

It is not clear how GBIF treats taxonomic synonyms. It does list names that are unambiguous synonyms of the name being searched, but whether records using such synonyms are provided with those being sought is not evident. Individual searches on *Agropyron spicatum*, *A. inerme*, and *Elytrigia spicata*, all of which are synonyms of *P. spicata*, brought up some records but, because there were less than 50 of them in total, it is doubtful that their exclusion affected the results obtained.

DISCUSSION

Clearly, the value of GBIF will increase as more specimen data become available. For this reason, we recommend that the International Triticeae Consortium actively seek the funding needed for making more data on the *Triticeae* available to GBIF. Every effort should be made to ensure that these funds allow for the cost of georeferencing specimens with clear locality information. Obviously, obtaining georeferenced data with a GPS unit at the time of collection is the most desirable situation. Unfortunately, even at their decreased cost, GPS units are still outside the budget of many taxonomists in countries where the *Triticeae* grow.

We have demonstrated the rather coarse depiction of the ecological limits of species that can

be obtained from retrospectively georeferenced specimens. Such information can be used to predict where else the species might be expected to grow. The accuracy of such predictions can, however, be greatly increased if both presence and absence data are available. Herbarium specimens represent verifiable presence data. Absence data are records for sites where a species was sought but not found.

There is, at present, no mechanism for storing such records, even if, as is rarely the case, they are made. Making and keeping such records is essential if we are to make better use of statistical modeling packets for spatial data (e.g. Statmod Zone – Garrard 2002) in looking for new locations, estimating the conditions needed to establish a new crop, or predict the spread of an introduced species. The Intermountain Herbarium has modified its database so that it can store such information, and GBIF is working on establishing standards for sharing such information (Speers, pers. comm., June 13, 2005). We shall urge other herbaria to start recording such information for it is a logical extension of the role of such institutions in maintaining a record of the world's plant biodiversity and its distribution through space and time.

CONCLUSION

In preparing this paper, we used ArcView, a program for which Utah State University has a license making it inexpensive for us to use. There are numerous free mapping and GIS programs available on the Web at OpenSource GIS (http://opensourcegis.org). We mention a few here to assist those interested in getting started. A search on the Web will locate many others. One program for individuals interested simply in producing distribution maps is MapMaker (http://www.mapmaker.com/). ArcExplorer is free software from ESRI that performs basic GIS functions (display, query, and retrieval). Free predictive modeling programs that can be used for presence-only data include DesktopGarp (Genetic Algorithm for Rule-set Production) available at http://www.lifemapper.org/desktopgarp/, and Biomapper (http://www2.unil.ch/biomapper/). Both DesktopGarp and Biomapper are based on the Ecological Niche Factor theory.

Acknowledgements. We thank Dr. Kathleen Capels for her editorial assistance in preparing this manuscript.

References

DeLorme (2004): StreetAtlas. DeLorme, Yarmouth, Maine.

ESRI (1998): ArcView 3.3. ESRI, Redlands, California.

Garrard C. (2002): Statmod: A tool for Interfacing ArcView® GIS with Statistical Software to Facilitate Predictive Ecological Modeling. [Master's Thesis.] Department of Biology, Utah State University, Logan, Utah.

Global Biodiversity Information Facility (2005): http://www.gbif.org. Accessed May 27, 2005.

U.S.D.A. Miscellaneous Publication No. 200. U.S. Government Printing Office, Washington, D.C.

Hitchcock C.L. (1960): Family Gramineae. In: Hitchcock C.L., Cronquist A., Ownbey M., Thompson J.W. (eds.): Vascular Plants of the Pacific Northwest, Part I: Vascular Cryptograms, Gymnosperms, and Monocotyledons. University of Washington Press, Seattle, Washington: 384–725.

Microsoft (2003): Excel. Redmond, Washington.

Oregon State University (2004): Spatial Climate Analysis Service. http://www.ocs.oregonstate.edu/prism/. Data accessed May, 2005.

Tutin T.G., Heywood V.H., Burges N.A., Moore D.M., Valentine D.H., Walters S.M., Webb D.A. (1980): Flora Europaea, Vol. 5. Cambridge University Press, Cambridge, U.K.