The Atlas of Living Australia's Spatial Portal

Lee Belbin
Atlas of Living Australia
leebelbin@blatantfabrications.com

***Abstract*—The Atlas of Living Australia is an AUS$65m Australian Government initiative to "To develop an authoritative, freely accessible, distributed and federated biodiversity data management system". The Atlas, led by CSIRO, partners with over 18 National, State and Territory agencies to deliver online, a wide range of biological and environmental information. The Atlas also supports nodes of, or links to the Global Biodiversity Information Facility, Catalogue of Life, Encyclopedia of Life, Biodiversity Heritage Library (BHL), Map of Life, Barcode of Life Data Systems (BOLD), Ocean Biogeographic Information System, Morphbank, the Taxonomic Database Working Group and other projects. Two years into the three year project, the Atlas delivers over 114,000 species and 22 million occurrence records, 200+ environmental layers, a range of spatial and annotation tools and citizen science support. The Atlas Spatial Portal, the focus of this paper, is a tool designed to support environmental research and management. The focus of the portal is species, areas, environmental layers, spatial analysis and data import/export.**

Keywords—GIS; spatial; Australia; biodiversity; tools; analysis; environment; management; conservation.

I. Introduction

The Atlas of Living Australia [1] is a Federal Government initiative to provide public access to the widest range of biodiversity-related data in the Australian region. Funding for the project came from the national Educational Investment Fund (AUS$30m), the National Collaborative Research Infrastructure (AUS$8m) and from in-kind contributions from partner agencies (AUS$26.5m).

The mission of the Atlas is "To develop an authoritative, freely accessible, distributed and federated biodiversity data management system". Fortunately, at the start of the project, Australia had useful infrastructure in place for sharing biological data. The Commonwealth Heads of Faunal Collections and Commonwealth Heads of Herbaria provided an existing structure for the sharing of faunal data through the Online Zoological Collections of Australian Museums - OZCAM [2] and Australia's Virtual Herbarium [3]. To this base, significant volumes of data have been added from Birds Australia [4] and various State and Territory collections. The coverage of the Atlas includes plants, animals and microorganisms, marine, terrestrial and limnetic species, native and non-native species and endemic and invasive species.

There are two closely related projects: the Terrestrial Ecosystem Research Network [5] and the Integrated Marine Observing System – iMOS [6]. From the Atlas perspective, TERN focuses on systematic bio-survey data while iMOS has focused on monitoring the marine physical/chemical environment. In addition to infrastructure related to species occurrence records, the Atlas also has significant projects related to citizen science, the Bioversity Heritage Library - BHL [7], Barcode of Life Data - BOLD [8], Morphbank [9] and Identify Life [10] that link with Atlas data.

II. User Needs Analysis

The Atlas commissioned an extensive user needs analysis report [11]. The key applications identified were species distribution analysis, species identification, site assessment, habitat management and planning, managing reference databases, public education, synecology and biosecurity. Three issues of particular significance were identified: resolving scientific names; integrating amateur observations and the management of sensitive data. As well as a feedback link on all Atlas pages, a comprehensive annotations service for species data was also required.

The target audience of the Spatial Portal is the scientific community and the key statement in the analysis was "Distribution analysis is the dominant task. The ability to retrieve spatial information will be essential – varying in time, varying in scale, with many different forms of content." The spatial priorities distilled to "Where does this species occur?" and "What species occur in this area?" We have extended both functions to include any taxonomic level and 13 different ways of defining 'area'. We have also placed a high priority on the ability to upload and download data. Web service access to all key Atlas functions was considered as a basic requirement and feedback can be submitted from any web page of the Atlas.

III. The Spatial Portal

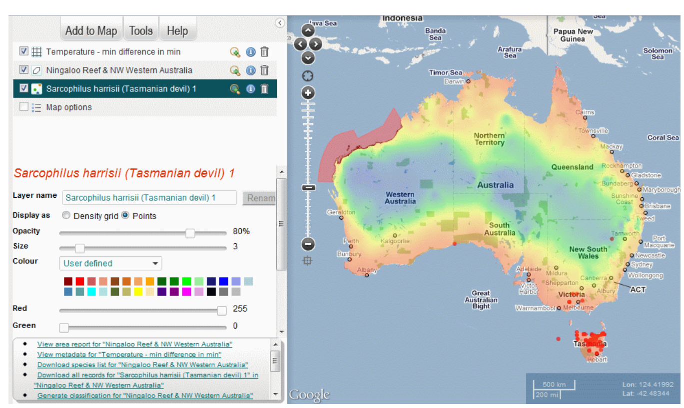

Figure 1. The Atlas of Living Australia's Spatial Portal.

The Spatial Portal of the Atlas (Fig. 1, http://spatial.ala.org.au) provides the main geocentric view of Atlas data. We have tried to align with Best Current Practice; building on what has been done effectively rather than reinvention. A review of over 30 existing geospatial portals provided useful criteria to address and a 'Google Maps/Earth look and feel' seemed to be assumed by most users.

To provide efficient functionality for the public and the research community was however no easy quest. We addressed this by trying to make simple and more complex functions equally intuitive. The 'Add to Map' tab provides rapid public display of taxa, areas and layers on the map while the Tools tab provide some form of analysis. Options for all functions are context sensitive.

We also attempt to lead the casual user to explore more advanced functions of the Spatial Portal. For example, the lower-left area of the Portal is used to provide hints of possible actions based on current mapped layers. For example, if a species has been mapped, the hint area provides links to species metadata, species records download, a scatterplot for the species and a spatial predictive model for the species. If an area is defined/mapped, hints include offering a checklist of all species in that area, a spatially predictive model of mapped taxa in that area or an environmental domains analysis for the area.

On the map window, there is only one function beyond the standard Google zoom, pan and 'zoom to my location functions': a layer interrogation button (a hover tool). On the command window, there are only three functions: 'Add to map'; 'Tools' and 'Help'. Taxa, areas and layers can be added to the map. Tools include species lists, sampling layer data at species locations, scatterplots, environmental domain generation and species spatial modelling.

Mapping data and most tool operations result in a map layer listed on the top left of the Spatial Portal. A set of icons for each layer provide on/off, layer type, zoom to layer extent, access to layer metadata and layer deletion. The center-left of the portal displays layer legends and some analysis results. The legend provides both the keys to the mapped layers and the ability to change the characteristics of the displayed layer. For example, points representing species locations can be sized, colored using a color palette or RGB slider bars, and the transparency adjusted. The legend also allows for colour faceting on various Darwin Core fields such as data provider, basis of the record and spatial uncertainty. This functionally will be extended to most Darwin Core fields.

The code base of the Atlas Spatial Portal came from the iMOS Ocean Portal [12]. Key components of the Spatial Portal include Java ZK code base, GeoServer, OpenLayers, GeoNetwork, Google API and OGC standards (predominantly WMS to date). The species occurrence data and their intersections with all the spatial layers are based on SOLR indexing of a Cassandra database. Most of the functions provided by Atlas components use RESTful JSON services. Atlas code and data are generally licensed under Creative Commons CC BY 3 [13].

A. Species

Taxa include point occurrence records (based largely on the Darwin Core standard [14], checklists (lists of species within a defined area) and 'expert distribution maps – polygons defining where a species should occur. The latter is a special case of the checklist. We hope to be able to include species tracking data in the next year.

An 'auto-complete' search strategy is used and scientific and common names with synonymies are supported. The auto-complete list also displays the type of record, taxonomic level and the number of occurrences.

Two additional taxa-related options are supported. A set of point coordinates or a set of Life Science Identifiers – LSIDs [15] can be imported in comma-separated variable (CSV) format for a portal session. The coordinate file currently supports three variables; a label, a longitude and a latitude. The option to import (CSV initially and then Darwin Core XML) up to 256 additional fields will be implemented in the near future and will support faceting of the records on all fields. The LSIDs can be of any taxonomic level entity held by the Atlas and can therefore be used to map and analyze assemblages. There is no QA performed on data uploaded for a session.

B. Areas

'Areas' correspond to objects held in our gazetteer, generated dynamically or uploaded to the Spatial Portal. The base used by the Atlas is an amalgamation of the 2010 Australian gazetteer [16], the Global Administrative Areas Database [17] and all of our named polygons (e.g., States and Territories) and classes (e.g., 'Forestry') from 'contextual' layers (see below). In all, there are 13 ways that an area can be defined by the Spatial Portal:

- Interact with the map (draw bounding box, polygon, point and radius or select area from contextual/polygonal layer);
- Search (radius centered on street address or a gazetteer polygon);
- Preset areas (Australia, world or current view);
- Upload (Shapefile, KML or WKT format) and
- Define environmental envelope.

The environmental envelope option is by far the most complex and powerful. This option uses slider bars on upper and lower bounds of one or more of the 150+ environmental layers to define an environmental combination and the corresponding area on the map. For example, you can identify where in Australia the mean annual temperature is between 10-12c and the precipitation of the driest quarter is between 150-250mm.

Sadly, the Australian Gazetteer only contains point locations. To enable users to determine what species are associated with a named gazetteer location, we have added an option to select a 1, 5, 10 or 20km radius around the points.

C. *Layers*

'Layers' are defined in a traditional GIS sense. In the Atlas, these can be terrestrial or marine and either 'environmental' or 'contextual'. Environmental layers are usually grids containing continuous values such as mean annual temperature. Contextual layers are usually polygonal in structure and contain class values. An example contextual layer would be 'Land Use' and a class within that layer could be 'Forestry'.

There is an obvious overlap between layers and areas. As noted above, we have included all the classes of contextual layers in our gazetteer even though some classes would be defined as named multiple polygons rather than single polygons. This strategy provides users with maximum flexibility in mapping features. For example, the polygons of the Land Use class 'Horticulture' can be mapped directly from the Areas option as well as the complete Land Use layer via the Layers option.

The 200+ layers available through the Spatial Portal (http://spatial.ala.org.au/layers) required two user-selection options. Layer selection can be done by an auto-complete with synonyms and tags supported. For example, precipitation and rainfall can be used synonymously. In some cases, we have included codes for well-known suits of layers. For example "Bio01" can be used as a shortcut for the "mean annual temperature" of the climatic layer suite of Hutchinson and Kesteven [18].

A two-level classification was also developed to guide new users through the layer library. The top level of this classification has the terms, area management, biodiversity, climate, distance, hydrology, marine, political, substrate, topography and vegetation.

IV. SPATIAL PORTAL TOOLS

Two workshops were run to determine what tools would be appropriate for the Atlas [19] and what data would be required to support these tools [20]. Criteria for evaluating tools included for example "accepted as best current practice", "wide applicability' and 'robustness'. Subsequently, the ability to import and export data was considered as a higher priority than adding novel tools. For example, technical users wanted to import species occurrence coordinates, append environmental and contextual values, export the records and then use the integrated data with their favorite desktop tools. A few exemplar tools were however required to demonstrate the wide range of applications that could leverage an extremely large volume of integrated biological, environmental and contextual data. We wanted to demonstrate a few of the possibilities.

As one of the reviewers correctly pointed out, the Spatial Portal is in part a "data discovery, integration and subsetting tool that produces customized subsets of data that scientists can use." Downloaded data comes with whatever (usually Darwin Core) attributes are supplied by the data provider, for example record identifiers. In most cases, LSIDs have been added for species if they do not exist in the original records. UUIDs have been generated for all defined areas and GIS layers.

Metadata for species data usually applies at the species collection level while metadata for 'GIS layers' is at the individual layer level. Analysis downloads also include a reference identifier that can be re-submitted to the Spatial Portal in subsequent sessions to re-create the analysis and the associated outputs and downloads.

All the tools can make use of any definition of area and in relevant cases, taxa. For example, after starting the Spatial Portal, the spatial prediction of taxa based on uploaded coordinates and available environmental layers over an area of choice (based on any one of the 13 options above) could be achieved in a few mouse clicks.

One of the highest priorities for the Atlas is addressing data quality or more appropriately the concept of 'fitness for use' [21]. For the biological data, this is the responsibility of the Data Management group within the Atlas of Living Australia project. The Atlas philosophy is to expose the data as received, enable annotations and only correct the 'bleeding obvious'. That said, the Atlas is in a good position to detect data issues and potential solutions and direct these to the data providers to address as needed. 'Fitness for use' would require a separate paper and is not addressed further here.

A. *Area Reports*

Like all other tools, the area can be predefined or generated on the fly using any of the 13 options above. The report lists the area (in square kilometers), the number of species, the number of occurrence records, the number of species polygons (from expert distributions and species checklists) and the number of related publications via BioStor [22]. From the report, the species lists and the list of full occurrence records can be downloaded as CSV-formatted files. The report also provides for direct mapping of all occurrences within the defined area.

B. *Species Lists*

A comprehensive list of all known species occurrences in the Atlas can be produced and downloaded as a CSV-formatted file for any defined area (a checklist); single or multiple polygons. The report lists the Family Name, Scientific Name, Common Name/s, Taxon rank, Life Science Identifier and Number of Occurrences. Filtering of sensitive species is performed according to the Atlas sensitive data service but no further analysis of the checklist is performed.

C. *Sampling*

This option is similar to the Species List except that all species occurrence records for any defined area are listed and downloadable in CSV format. Key Darwin Core fields [23] are included. Optionally, any combination of the 200+ layer values (environmental or contextual) can be appended to the occurrence records for download. This option also enables the appending of layer values to uploaded coordinates (which are treated as just another biological GIS layer), which can then be downloaded. Sampling is perceived as probably being the most useful tool for the key target audience (environmental scientists and managers). Data can be integrated, subset, downloaded and readily ingested into the tools of choice. Scripts are being written to simplify ingestion of the downloaded data into packages such as R [24].

D. Scatterplots

Checking for 'environmental outliers' among occurrence records was seen as one way of making effective use of the environmental layers to be found through the Spatial Portal. A scatterplot tool was however likely to bring a far wider range of benefits for examining the environmental conditions associated with species occurrences.

The scatterplot tool (Fig. 2) accepts any two taxa (a primary and a secondary) in any defined area and any two environmental layers. The scatterplot also identifies all possible environmental combinations within the defined geographic area. A grey-scale is used to display the spatial extent of the environmental combinations; darker cells represent small geographic areas while large areas are displayed as lighter cells.

The scatterplot tool is interactive: Selecting occurrence records on the scatterplot (environmental space) will select the occurrences on the map (geographic space). Occurrence records with missing values are also separately identifiable. Included and excluded subsets (with either including the missing value records) can be automatically generated from the scatterplot. Each subset creates a new mapped (bio-) layer within the Spatial Portal and can be used for any subsequent analysis and download.

In the near future, we will generalize the scatterplot tool to accommodate any pair-wise combination of environmental and contextual layers. For example, we see great benefit of being able to cross tabulate say land-use (contextual) and dynamic land cover (contextual) or mean annual temperature (environmental).

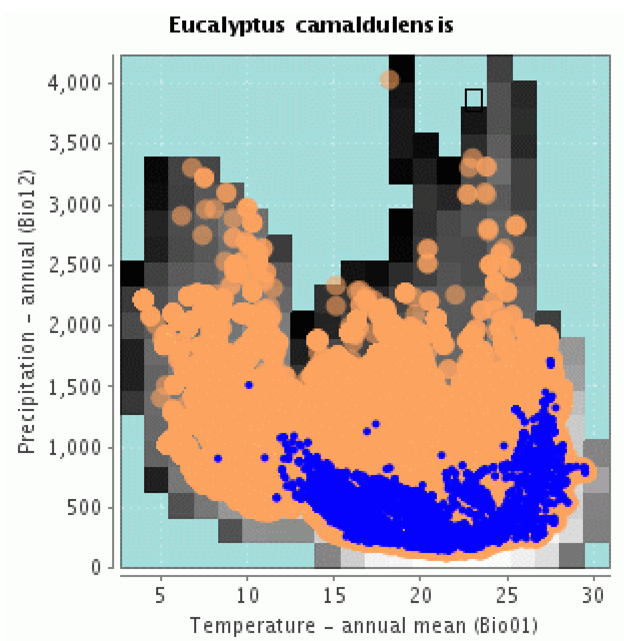

Figure 2. Scatterplot of *Eucalyptus camaldulensis* (primary – colored blue) with all *Eucalyptus* (secondary – brown) plotted against mean annual temperature and annual precipitation for contiental Australia. Greyscale envelope represents all possible terrestrial environmental combinations.

E. Classification

How do we make rational land management decisions where inadequate biological data is the norm? The use of environmental domains [25] may help. If species respond to environmental factors, it makes sense to use environment as a surrogate where adequate biodiversity data is lacking. Environmental domains result from a classification of multiple environmental layers.

The Spatial Portal contains a classification method [26] designed to generate environmental domains: areas of similar environmental properties based on multiple environmental layers (Fig. 3). One new environmental domains layer hopefully contains the most salient features of all submitted layers. The group colors are designed to represent group differences [27]. Fig. 3 provides a realistic ecosystem view of Australia based on only three environmental layers: annual precipitation; temperature annual minimum mean and a fertility scale based on lithology.

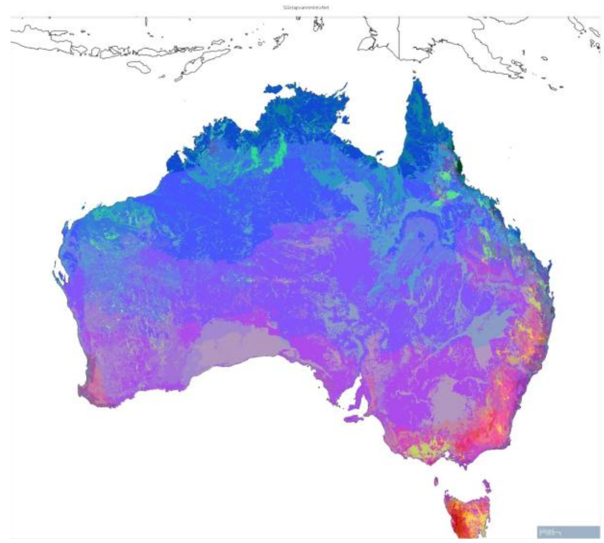

Figure 3. Spatial representation of the 51 group environmental domain classification of the Australian continent generated by the ALA Spatial Portal classification function using annual precipitation, temperature annual minimum mean and a fertility scale based on lithology.

F. Prediction

Spatial prediction models were seen as a stereotypic method demonstrating the value of integrated biological and environmental data. Such models provide environmental interpolation; displaying where species *could* occur and with what probability. Like any tool, it can be abused but that is not the focus of this paper.

MaxEnt (maximum entropy density estimation) [27] was recommended by the tools workshop [19] and is implemented in the Spatial Portal as an external package. Extensive testing suggested that realistic models could be generated from environmental layers that were known to constrain the distribution of the species in some way (Fig. 4).

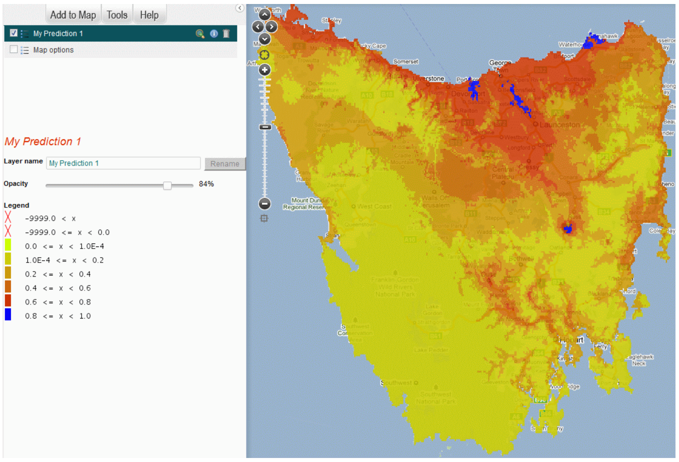

Figure 4. MaxEnt spatial prediction of Sarcophilus harrisii (Tasmanian Devil) using Aridity index - month max, Precipitation – seasonality and Temperature - annual max mean.

CONCLUSION

The Spatial Portal of the Atlas of Living Australia aims to address the user requirements of the research, environmental management, environmental consultant and environmental NGO communities for the provision of biological and environmental information. A free an open source (FOSS) strategy based on a Google Maps look and feel was foundational. Designing an interface that would support the casual user and the target audience was a challenge addressed by simplifying to key functions and context sensitive prompts.

The focus of the Portal is the integration of biological and environmental/contextual data. Integration of Australia's biological data is a significant end in itself, but linking these data with a wide range of environmental and contextual layers provides opportunities to support an extremely wide range of environmentally-related applications. Portal tools such as spatial modeling of species demonstrate the value of integration but the export (or import) of occurrence records with optional environmental and contextual values appended probably provides the greatest utility for the target audience. The current round of Government funding of the Atlas ends June 30, 2012. The project is based on open source principles and has established protocols with a wide range of organizations. This strategy aims at automating, minimizing and spreading the load of ongoing development and maintenance.

ACKNOWLEDGMENT

My appreciation goes to Donald Hobern, the Director of the Atlas of Living Australia for providing me with autonomy, a budget to achieve a significant outcome (that I leave others to judge), astute observations and endless encouragement. The Spatial Portal would not be possible without the enthusiasm, skill and dedication of the spatially scattered team: Adam Collins; Ajay Ranipeta; Angus MacAulay; Gavin Jackson and Brendan Ward. The comments of the three reviewers are also gratefully acknowledged. My thanks also go to Miles Nicholls and Bryn Kingsford whose tenacity in tracking down data and metadata will I hope be appreciated by many.

REFERENCES

[1] The Atlas of Living Australia: http://www.ala.org.au.

[2] Online Zoological Collections (OZCAM): http://www.ozcam.org.au/.

[3] The Australian Virtual Herbarium (AVH): (http://www.anbg.gov.au/chah/avh/avh.html.

[4] Birds Australia: http://www.birdsaustralia.com.au/.

[5] Terrestrial Ecosystem Research Network (TERN): http:/www.tern.org.au/.

[6] Integrated Marine Observing System (IMOS): http:/www.imos.org.au)

[7] The Biodiversity Heritage Library (BHL): http://www.biodiversitylibrary.org/.

[8] Barcode of Life Data (BOLD): http://www.boldsystems.org/.

[9] Morphbank: whttp://www.morphbank.net/.

[10] Identify Life: http:/www.identifylife.org/.

[11] The Atlas of Living Australia User Needs Analysis report http://www.ala.org.au/about/communications-centre/publications/user-needs-analysis-report/.

[12] iMOS Spatial Portal: http://imos.aodn.org.au/webportal/.

[13] Creative Commons CC BY 3.0 license: http://creativecommons.org/licenses/by/3.0/au/deed.en

[14] The TDWG Darwin Core standard: http://www.tdwg.org/activities/darwincore/.

[15] The TDWG Life Sciences Identifier Applicability Statement: http://www.tdwg.org/standards/150/.

[16] The Australian Gazetteer: https://www.ga.gov.au/products/servlet/controller?event=GEOCAT_DETAILS&catno=71110.

[17] Global Administrative Areas Database: http://www.gadm.org/.

[18] Hutchinson, M.F. and Kesteven, J.L., 1998. Monthly mean climatic surfaces for Australia. Unpublished. http://fennerschool.anu.edu.au/publications/software/creswww.pdf

[19] Flemons, P., Raymond, B., Brenton, P. and Belbin, L., 2010. Atlas of Living Australia report on the spatial analysis toolkit workshop. Unpublished. 2010. http://www.ala.org.au/wp-content/uploads/2010/09/Lee_ALA-Spatial-Ananysis-Tools-Workshop-Report-V3.pdf

[20] Flemons, P. and Belbin, L., 2010. Report on the Environmental data library workshop. Unpublished. http://www.ala.org.au/wp-content/uploads/2010/08/Lee-NEDL_Workshop-Report-Final.pdf

[21] GBIF position paper on Fitness for use: http://www.gbif.org/communications/news-and-events/showsingle/article/gbif-position-paper-enhancing-fitness-for-use-across-gbif/.

[22] Page, R.D.M., 2011. Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library", BMC Bioinformatics, 12, 187. doi:10.1186/1471-2105-12-187.

[23] TDWG Darwin Core standard terms: http://rs.tdwg.org/dwc/terms/.

[24] The R Project for Statistical Computing: http://www.r-project.org/.

[25] Nix, H.A., 1986. A biogeographic analysis of Australian elapid snakes. In: R Longmore (Editor), Atlas of Elapid Snakes of Australia., pp. 4-15. Australian Flora and Fauna Series Number 7. Australian Government Publishing Service, Canberra.

[26] Belbin, L., 1987. The use of non-hierarchical allocation methods for clustering large sets of data. Australian Computer Journal 19, 32-41.

[27] Belbin, L. Marshall, C. and Faith, D.P., 1983. Representing relationships by automatic assignment of colour. The Australian Computing Journal 15, 160-163.

[28] Phillips, S.J., Dudík, M. and Schapire, R., 2004. A maximum entropy approach to species distribution modeling. In: Carla E. Brodley (Editor), Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Canada, pp. 655-66.