Project no. 018340

Project acronym: EDIT

Project title: Toward the European Distributed Institute of Taxonomy

Instrument: Network of Excellence

Thematic Priority: Sub-Priority 1.1.6.3: "Global Change and Ecosystems"

C5.070 Report of converter webservice that transforms point occurrence data to distribution data (report 1)

Report 2 (C5.070: M41)

Due date of component: Month 34
Actual submission date: Month 35

Start date of project: 01/03/2006 Duration: 5 years

Organisation name of lead contractor for this component: 16 MIZPAN, 4 CSIC and 14 RMCA

FINAL

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

C5.070 Report of converter webservice that transforms point occurrence data to distribution data (report 1)

This report provides a summary of the procedures and steps to follow using converter webservices for transforming point occurrence data to distribution data.

Point data to regional distribution - Priorities in programming efforts should thus be concentrated on a set of known end-user needs in the domain of taxonomy related work, for practical purposes. Based on previously conducted surveys and reports following services have been selected as a priority to be used by taxonomists or users of taxonomic information:

Logical aggregation -qualitative generalization to a higher logical aggregation level. EDIT mapViewer currently prints user data based on *Genus* information. A logical aggregation procedure is thus followed by the system already. Any kind of logical aggregation can be applied if user has the necessary information, but the range of possibilities is so high that requires a very flexible interface and programming.

Spatial reference – already aggregated point data (aggregation might be at any level) should be referred to an area. In the area the points are referred to might be existent, such as administrative units, statistical divisions, climate types, geographical units, etc. or user-defined such as regular grids of squares based on different coordinate systems (UTM, AFE, local grids) or other regular grids such as triangles, hexagons, etc.(polygons) (*).

These kind of operations are foreseen in EDIT mapViewer as a measure of inventory facility, that is:

*"do I have enough data **in my selected spatial units** in order to apply possible future modeling?"*

Some current GBIF and ALA (Atlas of living Australia) tools can be of help here (see: http://alatools.pbwiki.com , see http://www.gbif.org).
A demo can be seen in the EDIT mapViewer, "Iberian Peninsula" context, "Spatial Completeness" prefiguring the full implementation.

Following maps can be found:

1) Number of records per spatial unit: gives an idea of the collecting efforts
2) Number of Genera per spatial unit: gives a (false) idea of the biodiversity
3) Number of Genera/Number of records per spatial unit: gives a more accurate idea of biodiversity
4) Uncertainty of Genera inventory: represents the mean effort dedicated to discover the last species found in each surface unit. This data can also be presented as a collector's curve (http://edit.csic.es/fitxers/curves/edit_curves.html).

All this data are visualized as choropleth maps that will let user chose in how many categories to distribute the data. Choropleth maps use relative data values –it means that the values of the represented data set are related to another base value (value per square km, percentage and so on)
More info on: http://edit.csic.es/web/docs/Demo_explanation_spatial.htm

Spatial Aggregation – spatially referenced data can be further aggregated if needed to a higher level of spatial aggregation. Since changing level of spatial aggregation requires two steps- relating input data to a chose reference unit (ie.parish) and then generalizing this unit to a higher level (ie.borough) it is worth to mention that the second steps is usually likely to be skipped-it is much more sensible to related the input data exactly to a chosen spatial unit.

In general points might be referred to any area type and further aggregation shouldn't be problematic-indeed this us the approach in the ABCD protocol and thus very much in the line with GBIF and other data connection initiatives. This is the best input data for a developer as it's relatively easy to handle points and convert them to areas using spatial sql or spatial overlay tools. Other conversions are given as examples of possible users needs and are based on the experience of the author. Although such cases are not common depending on the format of the source information of the specific end-use of the data conversions that may be considered as strange or un-common are asked. If the implementation of these cases is not too demanding, the suggestion is to offer them as additional services.

A very general algorithm for handling point to distribution conversion could look like as (just an demo sql script):

1. Qualitative aggregation of points to a needed level
   SELECT * FROM [InputPoints] WHERE [AggregationLevel] = Value; value might be i.e. genus name

2. Spatial referencing of points to a needed level of given division scheme
   SELECT [ReferenceAreaName], Count([InputPoints]) / [ReferenceAreaArea] as [Density] FROM [ReferenceArea] WHERE [ReferenceArea] Contains([InputPoints])

3. Visualize the data as a choropleth map

Distribution data sets are usually presented as choropleth maps. It is crucial to remember that choropleth maps should present relative values only. For absolute values diagrams are used (**).

Spatial operations – spatially referenced data may be further compared between data sets. it is possible to 'ask' spatial questions on point datasets (usually distance matrices and their variations or surface/grid interpolations) but spatial querying using vector data (areas lines and points) is much more common. Having one or more point datasets spatially aggregated it is possible to check many different spatial relationships between datasets for example:

- Show the areas that are occupied by the species A and B
- Show the areas that are occupied by the species A but not by species B or C
- What is most common land use category occupied by a given species

Conclusions and final remarks

As a final note, we can say that most of these operations could be developed as a GeoREST service, as the current EDIT geographic tools are. These services can then be used through an online tool, or as black-box (invisible to the user) background service for other tools. Uploading point data as CSV (a plain-text comma separated value file) should be implemented for working with big datasets. The current GeoREST point service (http://dev.e-taxonomy.eu/trac/wiki/MapRestServiceApi), is step in the good direction in this regard.

ANNEXES

Example of a logical aggregation

Level 3: Vineyard	Level 2: Permanent crop	Level 1: Agricultural Area
Mormoopidae *Pteronotus personatus*	Mormoopidae *Pteronotus*	Mormoopidae

Possible input data and types of conversion to area:

1. Points	1. Existing areas such as administrative units, geographical units, etc. 2. Regular grids based on coordinate systems (UTM, AFE, local grids) 3. Other regular grids
2. Areas	1. Areas might be aggregated to a higher level of given division scheme, i.e.– county → country or parish → county
3. UTM Squares	1. UTM squares might be aggregated to higher level of UTM division scheme, i.e.: 1km→10km or 1km→ 100km
4. AFE Squares	1. AFE Grids (the old and new one) have only one level of aggregation – square 50x50km
5. Other geometric grids	1. Areas might be aggregated to a higher level of given division scheme

A sample of spatial aggregation levels

parish	district	borough	county	country
UTM 1x1km	UTM 10x10km	UTM 25x25km	UTM 50x50km	UTM 100x100km

(*) As a side note, it's worth adding that sometimes a succession of different conversions might be needed by users – *conversions between reference systems, such as: Areas UTM, Areas / AFE, UTM AFE, AFE UTM, Local grid / UTM, etc.*

It is worth remembering that such conversions cause data blurring and might be used only for visual presentation of the data. Such conversions are irreversible – conversion from UTM to AFE and back to UTM will not produce a dataset identical with the input dataset. Such conversions are often relatively difficult tasks and are beyond the scope of mapViewer's functionalities, but end-users should be warned and highly recommended to always keep an original copy of the source data before using the services.

(**) Historically the presented values were referred to area of the unit only but with time cartographers started presenting values related to different parameters such as percentage. Presenting absolute values on a choropleth map is considered a bad cartography and should definitely be avoided.