Digital Databases and the Multiplicity of Biodiversity Views

Jorge Soberon
Department of Ecology and Evolutionary Biology and Museum of Natural History
University of Kansas

Terrestrial mammals numbers
Soberon & Ceballos, 2009
First, what is *biodiversity*?

- Simply put, the total manifestation of life on the planet. This is the “political” definition.
- The devil is in the details
- Biodiversity is not a simple concept, like, say temperature, that can be measured with thermometers.
What is biodiversity? II

- Biodiversity is a complex and badly defined concept.
- It makes reference to other complex and poorly defined concepts.
- And, alas, there are no biodiversitometers.

Somebody please invent Mr. Spock’s Tricorder.
Surrogates of Biodiversity

- Morpho
- Eco
- Phylo
- Geno
- Proteo
- Ethno
TaxoView

Pericopidae.

Biologia Centrali Americana
Fig. 3. Canonical variate analysis of the differentiation among sampling sites for *Auca coctei*. Color represents a shape variable which groups all individuals from the same site. *Dark red* points: La Serena (a); *dark green*: Collipulli; *dark blue*: Tumbes (c); lilac: Valle Elqui; *light green*: Dichato; *orange*: Concepción; *red*: Valparaiso; *calypso*: Los Lagos (b); *fluorescent green*: Loncoche. a1, b1, c1: female; a2, b2, c2: male.

EcoView

Full food webs of the Eastern Tropical Pacific (ETP)

And the Central North Pacific (CNP)

Hinke et al. 2004 Ecology and Society 9(1):10
PhyloView

ChemoView

A) Relative fitness vs. Scopolamine

B) Relative fitness vs. Hyoscyamine

C) Fitness vs. Number of trichomes

D) Fitness vs. Glucosinolates
EthnoView

*Datura stramonium*

*Libelus*, de la Cruz, 1552

*Quatro Libros*, Hernandez, 1652

*Mexican specimen*, Kew, 1874
Or, preferably, integrated views

Biodiverse, by Shawn Laffan. [http://code.google.com/p/biodiverse/]
Integrating knowledge domains

• This should be the future. Integrating different points of view.

• What is integrating?
  – Simplest: overlaying
  – More difficult: statistical relationships
  – Truly difficult: theoretical

• Whatever the meaning, to integrate one needs *Digitally Available Knowledge* (DAK, Sousa-Baena et al. 2013 Diversity & Distributions 20:369-381)
So, what data is available?

• At a global level, mostly occurrence data
• The data about an observation, or a specimen of:
  – What (name)
  – Where (lat, long)
  – When (date)
  – Says who (collector)
• Darwin Core, ABCD…

Images from the Smithsonian NMNH
Presence only data is abundant and useful (although not for children)
The Miridae (data from Weirauch et al.)
Alouatta palliata
INBio, Costa Rica
Antilocapra americana
Smithsonian NMNH
Mammal potential numbers since Interglacial (120,000 BP)

Purple ~ 80
Blue ~ 70
Green ~ 60
Yellow ~ 50
Red > 40
By integrating, the above can be...

- Dynamic maps of body size, ethnobotanic knowledge, secondary chemistry, interactions...
- Interacting, maybe dynamic perspectives on different knowledge domains
- The name is a link to attributes
- And the coordinates are a link to where
- These two are powerful links
There are already quite a few things to integrate with

- OPEN Tree of Life
- MORPHOBANK: Homology of Phenotypes Over the Web
- TRY: Plant Trait Database
- COMPADRE: Plant Matrix Database
- COMADRE: Animal Matrix Database
However...

- Digitizing is expensive

<table>
<thead>
<tr>
<th>Digistreet</th>
<th>Target</th>
<th>Realized</th>
<th>Average Cost per Object in Euros (including overhead)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herbarium sheets—Leiden</td>
<td>3,800,000</td>
<td>450,000 (3,500,000 scanned)</td>
<td>1.29</td>
</tr>
<tr>
<td>Microscopic slides</td>
<td>900,000</td>
<td>350,000 (590,000 scanned)</td>
<td>1.57</td>
</tr>
<tr>
<td>2D objects (books, journals, etc.)</td>
<td>900,000*</td>
<td>600,000*</td>
<td>1.87</td>
</tr>
<tr>
<td>Entomology</td>
<td>850,000</td>
<td>600,000</td>
<td>1.51</td>
</tr>
<tr>
<td>Herbarium sheets—Wageningen</td>
<td>800,000</td>
<td>800,000</td>
<td>1.47</td>
</tr>
<tr>
<td>Mollusks</td>
<td>510,000</td>
<td>640,000</td>
<td>1.37</td>
</tr>
<tr>
<td>Vertebrates dry</td>
<td>325,000</td>
<td>180,000</td>
<td>2.37</td>
</tr>
<tr>
<td>Geological specimens</td>
<td>220,000</td>
<td>70,000</td>
<td>1.90</td>
</tr>
<tr>
<td>Wood samples</td>
<td>125,000</td>
<td>125,000</td>
<td>1.27</td>
</tr>
<tr>
<td>Specimens preserved on alcohol</td>
<td>100,000</td>
<td>70,000</td>
<td>4.65</td>
</tr>
<tr>
<td>Total</td>
<td>8,530,000</td>
<td>3,885,000</td>
<td></td>
</tr>
</tbody>
</table>

*Number of pages.

Heerlien et al. 2015 ACM J. Comput Cult. Herit. 8:1-11
Which means that it pays to digitize first the large collections

Log 10 of cost per specimens of digitizing collections, as a function of their size. Data from CONABIO, Mexico. Cost per specimen $1-$10 USD
And not only expensive in $$$

- It requires time
- Personnel
- Space
- A long term commitment (databases are never finished)

- It will never be done unless the primary providers (collections) find uses.
- This means large-scale digitizing beyond labels will not take place unless it is in the direct interest of curators and systematists
Bute the benefits will be huge: Overlay

Figure 1. A screen-capture of the QGIS Lifemapper plugin showing a sites-based output. Site similarity plot for amphibians of the Philippines, showing highlands in Luzon (A, in yellow).

The value of the mean proportional range size (B), the emerged relief during the glacial maximum (C), and the "brushed" Luzon cells in the scatterplot of richness vs. mean proportional range size (D).

Figure 2. A species-based screen-capture of the QGIS plugin showing part of a mammal phylogeny (A) connected to a 800+ mammals of Africa PAM, with a map of species richness (B), mean proportional species-diversity (C), mean nearest taxon distance (D) and a scatterplot of range-size vs. mean phylogenetic distance (E) with the "brushed" species highlighted (yellow species in (A) and (E)).
Figure 1  Schematic arrangement of data matrices $I$, $E$, $B$, $P$ and $T$ and the three association ($D$, $\beta_E$, $\beta_B$) matrices discussed in the text.
Theory

• A real challenge.
• Do we have a “theory of biodiversity”?
• Of course we have a Theory of Evolution, but this is too general.
• A theory of biodiversity may emerge from the integration of those different points of view.
Maybe this should be the future

• Creating DAK in different domains.
• Morpho
• Eco
• Phylo
• Creating the concepts and tools to integrate them
• Many doubt we can do it

• Will it be a case of “ignoramus et ignorabimus” (du Bois-Reymond)
• Or rather one of: “Wir müssen wissen — wir werden wissen!” (Hilbert)
Thanks to...

- Organizers of the symposium
- Jeff Cavner, A. Christen, H. Arita, P. Rodriguez, A. Lira, F. Villalobos, coworkers in the biodiversity perspectives stuff
- A. T. Peterson, for endless conversations on this.
- And the money folks