

Entangled Bank: connecting ecological data with a semantic web framework and simple metadata

DAVID M. KIDD¹, David Orme², Georgina M. Mace^{1,2}, Tim. N. Coulson^{1,2}, Andy Purvis^{1,2}, Ian Owens^{1,2}, Sarah Knight¹, Thomas Ezard², Robin Freeman³, Martin Calsyn³, Eric Hellmich³ and Rich Williams³

¹NERC Centre for Population Biology and ²Division of Biology, Imperial College London, UK, ³Computational Ecology and Environmental Science Group, Microsoft Research, Cambridge UK

Integrating different types of ecological data provides the ability to address novel and important scientific questions, however such data are both highly heterogeneous and widely distributed across institutions and between individual researchers. Even once located, synthesis is made more problematic due to syntactic and semantic ambiguity and restricted metadata.

These difficulties mean that synthetic ecological science is enormously labour intensive, as data integration is done by hand. New technological solutions are required to overcome these obstacles and facilitate synthetic

research spanning multiple levels of biological organization from individual organisms to ecosystems whilst incorporating geographic and temporal variation. Such studies are often vital to developing the understanding of the natural world necessary for dealing with increasing anthropogenic pressures.

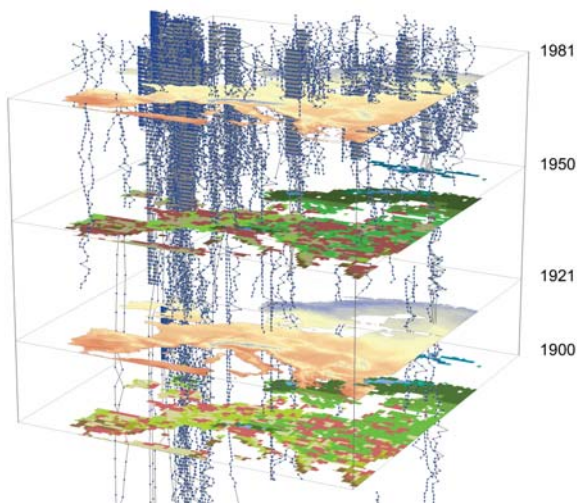
Entangled Bank is a joint project between the Centre for Population Biology and Microsoft Research that is developing informatics solutions to these problems. This is underpinned by a flexible data storage solution, based on a Resource Description Framework (RDF) data model with extended attributes for data con-

fidence, ownership and versioning. In order to promote use of the system, data storage will be coupled to a suite of web-based tools that will support the input, identification, management, processing, analysis and visualization of data and facilitate the maintenance of a core metadata description permitting the synthesis of data in a spatio-temporal framework.

The case studies below use a range of datasets hosted by the Centre for Population Biology to demonstrate the benefits of applying such a system in enabling complex data integration problems.

Responses of population dynamics to change in the environment.

Linking reconstructions of climatic and land use change over time with spatial information on population trends from the Global Population Dynamics Database (GPDD) can reveal which processes affect population abundance and variation.

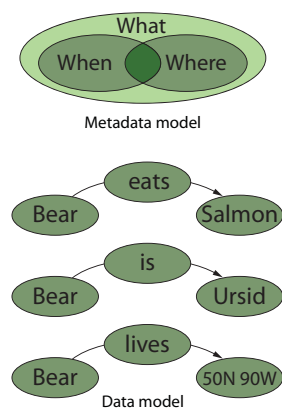


2900 abundance time-series from the Global Population Dynamics Database¹. Series are standardized with populations increasing to the east (right) and decreasing to the west (left). The time series intersect layers of annual mean temperature in 1921 and 1981² and land use in 1900 and 1950³.

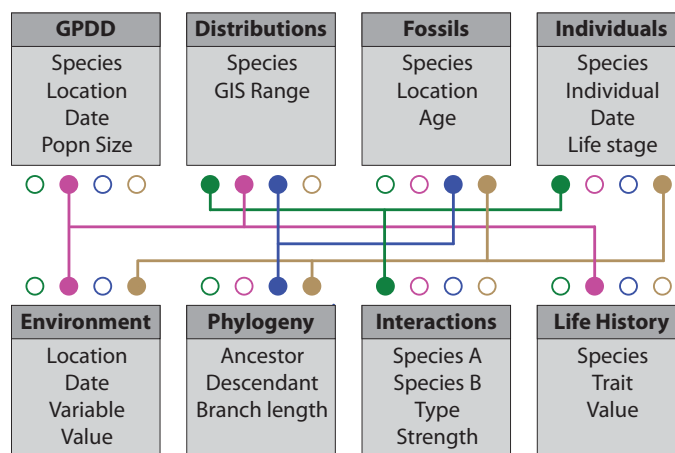
¹ NERC Centre for Population Biology, Imperial College (1999) The Global Population Dynamics Database. <http://www.swi.ac.uk/cpb/gpdd.htm>. ² Mitchell, T. D., Carter, T. R., Jones, P. D., Hulme, M. & New, M. (2004) A Comprehensive Set of High-Resolution Grids of Monthly Climate for Europe and the Globe: The Observed Record (1901-2000) and 16 Scenarios (2001-2100) (Tyndall Centre for Climate Change Res., Norwich, UK.), Working Paper 55. ³ Global Historical Land Cover and Land Use Estimates (1700-1990). Klein Goldewijk, K., 2001. Estimating global land use change over the past 300 years: The HYDE database, Global Biogeochemical Cycles 15(2): 417-433.

Conceptual Model

"A little semantics goes a long way" James A. Hendler



The core data are recorded as RDF triples, within a data store. A core subset of Ecological Markup Language provides heritable metadata at the document and triple level that provides the ability to match location, time and observation types across datasets.

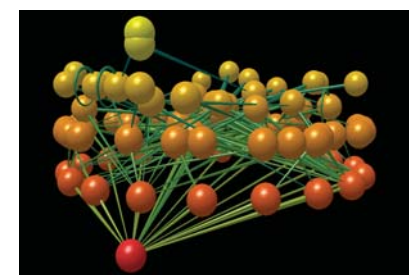


Interaction networks

Data gathered at Silwood Park on local plant species diversity, networks of species interactions, the distribution and behaviour of individual vertebrates can be combined with local environmental data.



Aerial image of Silwood Park, Ascot, UK with superimposed layers showing total plant species richness in 100m quadrats (scale at right) and estimated home ranges of nesting blue tits (*Parus caeruleus*). Nests with more solid ranges successfully raised chicks in 2005.



Food web of species associated with Broom (*Cytisus scoparius*) at sites in Silwood Park (Memmott et al. 2000, J. Anim Ecol. 69: 1-15).

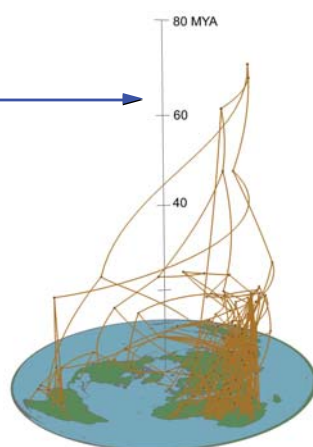
Reconstructing ancestral ranges through time

Reconstructing ancestral ranges requires the synthesis of phylogenies with current distributions, fossil evidence and palaeo-environmental reconstructions to create spatially referenced geophylogenies.

Paleogene (45 mya) and Neogene (12 mya) distributions of fossils (<http://paleodb.org>)



Geophylogeny depicting pattern of diversification in space and time.

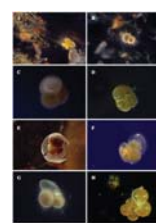


Supertree of extant Artiodactyla species (Bininda-Emonds et al 2007 Nature 446)

Extant Artiodactyla species ranges (Global Mammal Assessment)

Patterns of speciation in the planktonic Foraminifera

An abundant fossil record, combined with paleoclimate reconstructions and phylogenies, permit detailed study of temporal patterns in the rates of morphological evolution, speciation and extinction.



Extant (left) and fossil (*Turborotalia*, below) foraminifera.



Paleoclimatic reconstruction of oxygen isotope ratios (above: Zachos et al 2001, Science 292:686) and phylogenetic reconstruction of planktonic Foraminifera (right: Pearson, 1992, Paleobiology 18:115).

