All the evidence in one place

Olga Lyashevska
olyashevska01@qub.ac.uk

Abstract

Biodiversity is a multi-dimensional concept that is represented by a large variety of measures. This complexity and lack of consistency limits the development of a coherent scientific understanding of biodiversity and how properties, such as ecosystem services, may depend on it. Here, I demonstrate that the formal discipline of creating a relational database (RDB) for information about biodiversity and its measures, is a useful tool in organising such knowledge into coherent sense. Following steps of the logical database design and data normalization to build a RDB, results in a formal definition of biodiversity within a well defined concept structure; mapping rules between the concepts of biodiversity and entities of RDB and a consistent information structure - all in one place. I show how this is then used to support evidence-based objective statements about biodiversity.

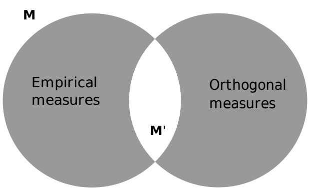
Concept definition

A database of measures of biodiversity will be constructed. It will be used to answer the following question:
'To what extent does existing knowledge describe the information content (=fundamental biodiversity) of biological systems?'

Fundamental biodiversity is the set of differences in measures *M* among a set of biological systems *(i,j)* whose component measures are strictly orthogonal.

Measure $M_{DL} \equiv (D|L)$ is a scalar combination of one descriptor D at one level L specifying a component of biodiversity. The matrix measure **M** is the set of all possible combinations $M_{DL}(\forall D, L)$.

N.B. M_{DL} can be null, so that **M** may be a sparse matrix.



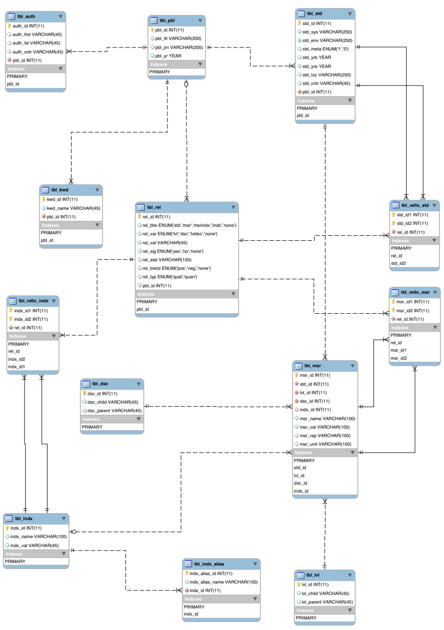
Empirical measures are measures collected from the literature;

Orthogonal measures are hypothetically possible measures that fully describe biodiversity;

Intersect M' is a description of biodiversity covered by literature, other sectors are unnecessary or missing;

Database specification

A relational database of measures of biodiversity `biodivDB` is running on a remote server. It can be accessed via both command line and phpMyAdmin. It contains 13 tables; the storage engine is InnoDB. To facilitate analysis of database it can be connected with R using RMySQL package.



The database will help to reveal patterns among biological systems and imply the minimum set of orthogonal measures.

Data dictionary

Table name	Description	Table name	Description
tbl_auth	author	tbl_pbl	publication
tbl_std	study	tbl_kwd	keyword
tbl_rel	relationship	tbl_tbl_relto_std	related to study
tbl_tbl_relto_indx	related to index	tbl_relto_msr	related to measure
tbl_dsc	descriptor	tbl_msr	measure
tbl_indx	index	tbl_indx_alias	index alias
tbl_lvl	level		

Cartesian Product $M_{DL} \equiv (D|L)$

Cartesian product is a way to extract matrix elements of **M**, so a view of cartesian join will be created. Every row of table `tbl_lvl` is joined with every row of `tbl_dsc`. This join produces a matrix of elements $(D|L)$, that is a matrix of all hypothetically possible measures made from descriptors *D* and a given level *L*.

```
CREATE VIEW tbl_cart1 AS SELECT lvl_child, dsc_child
FROM tbl_lvl, tbl_dsc;
```

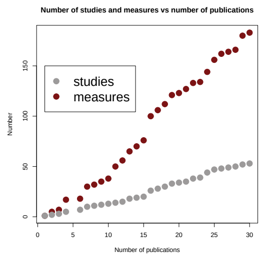
It generates 598 rows where each row is an element of the M_{DL} matrix. To compare generated measures with measures actually present in `tbl_msr` I run the following query:

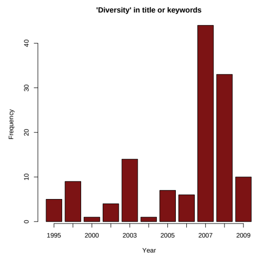
```
CREATE VIEW tbl_cart2
AS SELECT distinct dsc_child, lvl_child, msr_name
FROM tbl_cart1 LEFT JOIN tbl_msr
ON cart1.dsc_id=tbl_msr.dsc_id AND cart1.lvl_id=tbl_msr.lvl_id
WHERE msr_name IS NOT NULL;
```

It generates 45 rows, which are in fact distinct measures of biodiversity presented in the database. Analogically I can find a set of values where `msr_name IS NULL`. The difference between the two sets will indicate measures that are $M_{DL} = \emptyset$. 518 rows or 92% of all hypothetically possible measures are missing.

Preliminary Results

- database contains 30 publications, 53 studies (31 distinct) and 189 (43 distinct) measures of biodiversity;
- word 'diversity' appears 134 times in title or keywords in the period from 1995 to 2009.
- number of measures increases at a rate greater than the number of studies with an increase in a number of publications;





- 103 measures provide a value of measure, corresponding to 15 levels and 17 descriptors;
- 95 measures out of 189 are used to create an index (49 distinct pairs);
- species richness and species abundance are the most frequently used measures (46 and 38 respectively);
- systems that are mostly measured are grassland (34), insects (33), and macrobenthos (22);
- most of the relationships established are on measures level (48);

Further Work

At the point when an additional publication will not produce any new empirical measures, it is said that the saturation point has been reached. When saturation point is identified I will quantify the Venn diagram. If the rate of growth of publication/measure ratio of biodiversity is known, the number of measures needed to fully describe the information content of fundamental biodiversity can be predicted with greater confidence.

Acknowledgements

I am grateful to Dr Keith Farnsworth for his supervision and EPA in Ireland for funding.

Queen's University Belfast
Biological Sciences School
Queen's University Belfast