The Potential for Online Recording Metrics: Recentness of Data

Introduction

This paper is a brief discussion on metrics for measuring the success of online recording. It focusses in particular on how recent the records are that come through an OR system.

Discussion

There are now many systems with which wildlife recorders can get their records entered into databases and shared with the wider conservation community. With this is mind it is worth thinking about how the assumptions and assertions around these systems can be tested and measured. One way of doing this might be to analyse datasets and databases numerically and look for patterns.

A couple of months ago, I was sent a selection of figures for the numbers of records entered into an LRC's database over the last decade. These figures were broken down by the year the records were taken to show the number of records for each year within a twelve month data input period, as in the example below.

Table 1: A small subset of the data sent by LRC

	Numbers of records inputted in a given year		
Years records were taken (subset)	**2009**	**2010**	**2011**
2011	NA	NA	28072
2010	NA	17907	19886
2009	11191	7321	8940
2008	4947	5121	3920
2007	2731	5443	5309
2006	4320	4048	1304
2005	1440	4010	1594
2004	2319	7267	1427
2003	1753	7586	2241
2002	3341	5397	1183
2001	4615	2240	28072
Total for all years	63990	122715	135247

These figures presented the opportunity to explore the idea of analysing LRC datasets with regard to the recentness of the data, i.e. how quickly the records are reaching an LRC database. In an ideal scenario the figures could be used to track change over time, which could in turn be used to estimate a figures for the future and therefore set targets.

LRC databases (and possibly all databases produced from public submission of records) are very variable, reflecting the multitude of different surveys and ad hoc records that constitute such databases. This is reflected by the fact that trying to

find statistical correlations in the data provided by the LRC was impossible. Only very weak ones could be found which are not worth pursuing.

However, that doesn't mean that there is nothing to be gained from examining the figures in different ways. It just means that looking for trends and using them for prediction could be very difficult. There is some analysis that can at least pose some interesting questions, such as the following example.

One measure of the recentness of records in a database could be the proportion of records that are entered into the database in the same year as they were taken (for example, a white letter hairstreak was recorded by a volunteer on 28th July 2012 and was entered into the database on 19th September 2012). Some figures of this kind are presented below, using data from the LRC database.

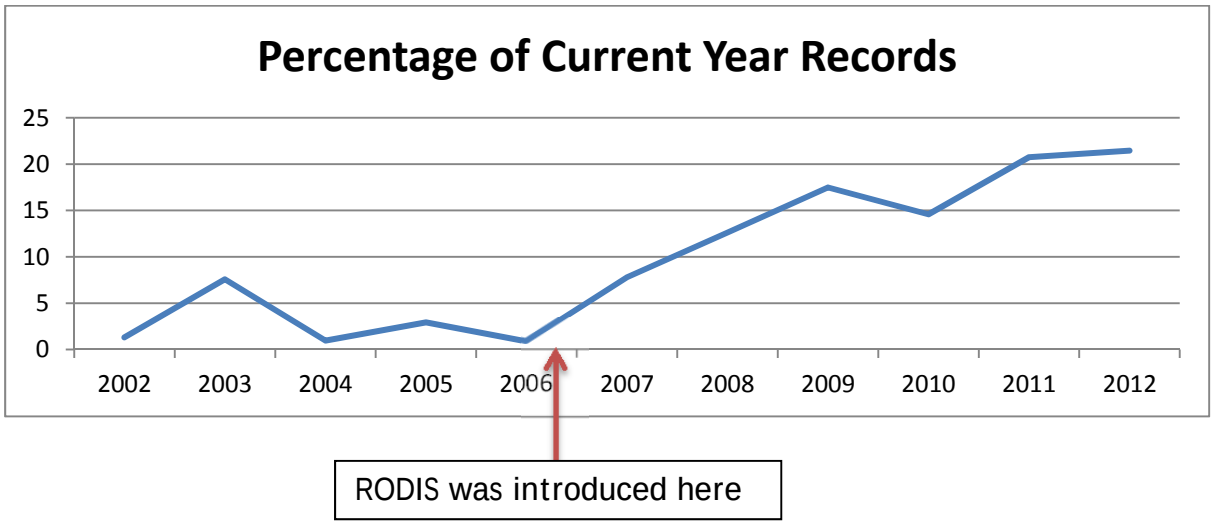
Table 2: Proportion of records entered into the database in the same year as they were taken

	Total records inputted	**Of which were for the current year**	**Proportion (%)**
2002	308918	4052	1.31
2003	155858	11811	7.58
2004	313500	3001	0.96
2005	64417	1884	2.92
2006	174120	1533	0.88
2007	234947	18315	7.80
2008	95370	12024	12.61
2009	63990	11191	17.49
2010	122715	17907	14.59
2011	135247	28072	20.76
2012	102458	21979	21.45

There is one obvious problem in looking at the figures in this way, and that is that it is bound by calendar year and does not therefore take into account the fact that records for say November and December could be inputted into the database in the early months of the following year, and still be considered to have been inputted recently. However, we know, and can prove, that most recording is done during the summer months, so it is not unreasonable to think that a good measure of the recentness of a database is the proportion of records that are inputted into it in the same year as they were taken. Future analysis could look at how many records were inputted within twelve months of being taken, and even calculate an average time from taking the record to inputting it into a database and seeing how this changes year on year.

The figures above are presented in the chart below.

Chart: Proportion of records entered into the database in the same year as they were taken



The chart suggests an almost continual increase in the proportion of records that are inputted in the same year during the second half of the ten year data set. This would seem to coincide with the introduction of the Record Online Data Input System (RODIS)[1] towards the end of 2006. However, before any conclusions on whether RODIS is causing records to enter the LRC database more speedily, further analysis would need to be carried out. The figures here should only be used as a potential indicator that a trend may exist and should inform more in-depth and robust indicators.

Conclusions

This brief exercise has given an indication of some of the figures that are available on wildlife recording databases and data input systems, and how analysis of these figures could be used to produce useful metrics. This is a flavour of what may be possible, but there will need to be further investigation if a full set of metrics is to be produced.

For example, what other effects might online recording have on databases? It could, perhaps, open up wildlife recording to a wider number of people and even a new generation who are very computer savvy. Most systems will have a user registration aspect, and it should therefore be possible to count the number of people using the system. This could become a more informative measure when compared to the amount of effort that is spent trying to train new users on the system. In the past there have been several LRC led training exercises, resourced

[1] This is an online recording system

by the NBN Trust, which have provided a number of people with access to online recording training. And so a useful metric might be to see how many active OR users this training generates.

Other aspects, for which there is no time available to discuss here, but might be worthy of consideration in future are:

- The volume and speed of progression of verified records.
- The proportion of records rejected by verifiers (because of the validation aspects of OR systems).
- The precision records (due to online mapping technology).