3.2. Information and data management

Marika Maxwell[1], Adrian Wayne[1], David McKenzie[2], Peter Orell[1], Sheryn Prior[1], Ian Wilson[1], Julia Wayne[1]

[1]Department of Environment and Conservation
[2]Private Consultant

Abstract

The use of existing data relating to fauna in the Upper Warren region to assist in the diagnosis of recent woylie declines was substantially compromised by the problems associated with the data being located on multiple isolated databases, maintained by multiple custodians and in various formats. In response, two closely-linked databases were created to manage the Woylie Conservation Research Project (WCRP) requirements; 1) Manjimup Fauna File (MFF) - an aggregation of existing fauna trapping data common to most datasets from the Upper Warren, and 2) Woylie PCS database to manage the detailed data unique and specific to the WCRP.

The principal datasets that were aggregated to create the MFF included the previously isolated databases associated with the Kingston Project, long-term monitoring of Perup, Bushrangers, *Western Shield*, DEC Fauna Management Course, and Donnelly District fauna activities. The aggregation resulted in 39,125 fauna records being incorporated into the MFF with the 12,000 existing fauna records. Over 25,000 woylie records spanning more than 30 years have subsequently been aggregated and made available for analyses that will assist in the diagnosis of the recent woylie declines.

Woylie PCS is the second, linked database and includes survival and mortality data from radio-collared woylies, and other research components including predators, resources, and disease. The centralisation of all woylie data into the two databases has meant that tracking and interrogation of data has been made much simpler. Once all disease data has been entered from respective external collaborators it will also allow for more complex analytical diagnosis between the WCRP components.

The methodology of the database developments are discussed as well as consideration of some of the challenges associated with this exercise. These achievements are a case-model and may assist in the development toward a Corporate-wide aggregation of related datasets.

3.2.1. Introduction

Existing data relating to fauna in the Upper Warren has historically been located on multiple isolated databases, maintained by multiple custodians and in various formats. Furthermore, the currency and the extent of quality control and data validation of these databases was highly variable. Among the many shortcomings of this, was that once a problem was recognized with woylie populations rapidly declining in some places it was not possible to readily interrogate existing data to develop a comprehensive and rapid regional assessment of the extent and nature of the woylie declines.

Some datasets took months to be brought up-to-date, clearing backlogs of unentered data extending into multiple years, notwithstanding the need for data validation to improve the quality and accuracy of the data. Many of the same animal individuals were recorded in multiple independent databases, each with a unique identification and data record reference system. Similarly there were differences in the format and management of associated data including what data was collected (sex, breeding and biometrics), units of measurement (mm or cm, g or kg, etc) and quality control measures at point of data entry. As a consequence, it remained impossible to get a comprehensive record of animal individuals of interest. Until these and related issues were resolved it remained impossible to satisfactorily investigate temporal, spatial and demographic differences and changes associated with woylie decline at a regional scale.

The very large amount of DEC fauna data collected over more than 33 years within the Upper Warren, if organized into a manageable form, would serve as an extremely valuable resource that

could assist substantially in the development of a reasonable understanding of the nature of the declines and provide potentially critically important evidence that could assist in the identification of the cause(s) of the woylie declines. This resource would be equally as valuable more generally, for the development of an understanding of regional fauna that would directly assist their conservation and management.

Therefore, an objective of the Woylie Conservation Research Project (WCRP) was to aggregate, as much as practically possible, the main relevant fauna-trapping datasets from the Upper Warren region into one standardised and centralised database. In addition to fulfilling the needs of the WCRP, the aim was to provide a resource that would manage all future fauna-trapping data from the region, and standardise data collection and management wherever possible and appropriate, while providing flexibility to accommodate specific requirements of independent projects. In this regard, this exercise and its product were considered a model and opportunity for the development toward a Corporate-wide tool, given that the same issues are common more broadly.

This exercise needed to be completed within a timely manner so that the aggregated data could be available for meta-analysis.

3.2.2. Methods

For efficiency, practicality and timely reorganisation of the fauna-related data, it was parsimoniously categorised into two;

- Data from fauna common to most of the principle datasets. This included trapping data and the standard biometric data of captured individuals. The principle datasets to be aggregated would include all of the 11 key transects involved in the Upper Warren Fauna Monitoring component of the WCRP. The datasets were;
  a) 'Fauna File' (1999 - 2007) – containing *Western Shield*, Fauna Management Course and other Donnelly District fauna monitoring data (Nature Conservation Division Custodians – Peter Orell, Ian Wilson/Julia Wayne)
  b) 'Kingston Project' (1994 – 2005) – containing terrestrial vertebrate data related to investigations into the impacts of timber harvesting (Science Division Custodian – Adrian Wayne)
  c) Long-term Perup fauna monitoring (1974 - 1999)– conducted by Per Christensen and Neil Burrows.
  …d) 'Bushrangers' (2000 – 2007)– a modified continuation of the long-term monitoring by Christensen and Burrows (Science Division Custodians – Bruce Ward and Graeme Liddelow)
- Data specific and unique to the WCRP, particularly the Population Comparison Study. This included data from the detailed examination of captured individuals (e.g. field health checks), samples collected from animals for associated research (e.g. disease, resources, genetics) and the results from subsequent tests on these samples, as well as data generated as part of the survival and mortality study of radio-collared woylies, predator sandpad surveys, and resources investigations.

On the basis of these classifications two distinct, but linked, databases were developed to manage these data respectively;

- Manjimup Fauna File (MFF)
- Woylie PCS

3.2.2.1. Manjimup Fauna File (MFF)

Fauna File is an existing database that was initially developed by Peter Orell in 1996 (Version 1.1 operational in 1997) for the original purpose of storing *Western Shield* data. This database grew to meet the requirements for all districts fauna trapping information and contains trapping and demographic information, as well as other fauna survey information including spotlighting and opportunistic sighting data). Fauna File has recently been revamped by Peter Orell and David McKenzie to help meet the requirements of this very large data set. Fauna File was chosen as the basis for aggregating all other data sets to create a Manjimup version of Fauna File (i.e. MFF). Donnelly District remains the custodian of the MFF.

The Woylie PCS is a database created specifically for the WCRP and contains all data for the Disease, Predator, Resource and Survivorship components of the project.

Aggregation of databases to create MFF

1. Datasets were itemised and prioritised for aggregation into MFF.
2. Datasets were located and the respective custodians enlisted for help in the aggregation process. The data was in a combination of electronic and hard copy format.
3. Standardisation of the data for importation. This involved a considerable amount of restructuring of the data into compatible fields and formatting of the data to meet the strict requirements of the database fields. Interpretation of the comments was required to match the data to the specific fields within MFF, for example, categorising reproduction comments into respective fields and matching to codes where possible. Editing measuring units. (Responsible - Peter Orell, data custodians, WCRP personnel)
4. Creation of new data fields within MFF to allow for the different information collected during various surveys. “Custom fields” were created and this allowed for the multiple fields of data within the constraints of the “Trapping Table”. (Responsible - David McKenzie and Peter Orell)
5. Creation of new fields in MFF to contain the meta-data for each of the datasets aggregated and for existing datasets. These included fields to contain information on data custodianship, background information to the custom fields, and comments on how and when the data was last updated. Provision is also made for rating each data record, one to five – nil to high level of validation, and to comment on the rating level for each data set. (Responsible - David McKenzie and Peter Orell)
6. Obtaining the original or ‘parent’ data from the respective custodians to assist with the data conversions and aggregations. (Responsible –WCRP personnel)
7. Follow a strict step by step process in MFF before trapping data can be entered. i.e. location details followed by date and personnel. Considerable effort was made to attain this information. (Responsible – WCRP personnel)
8. Importation by David McKenzie of datasets into Fauna File.
9. Data entered directly into MFF from hard copy data sets where electronic versions were not available or not suitable (Responsible – WCRP personnel and Verna Tunsell).

3.2.2.2. Woylie PCS database

Primary Objective

To create a database that would centralise all data pertinent to the WCRP.

Woylie PCS database requirements

- Fully functional during operational period of WCRP
- Efficient data entry with quality control measures at time of entry
- Security of data
- Autonomous and flexible
- User friendly
- Collate all results from the variety of external sources into the Woylie PCS database (e.g. test results from disease investigations).
- Allow for combined interrogations from different sections of the study
- Linked to Manjimup Fauna File (MFF)

Creation of Woylie PCS

1. Personnel were responsible for designing the structure and formatting of the database sections relevant to their own components of the study.
2. Information was then provided to David McKenzie for compilation and creation of the database.

3. The sections of the database were then interrogated and tested with real data. Feedback was provided to David McKenzie for any alterations and 'debugging'.
4. Over time modifications were made to the database in response to changes in data collection methods, especially with respect to disease sampling.
5. Security and quality control: The database was stored on the local network and read/write access provided to core personnel involved in the project only. Read only access was provided to other personnel.

The Woylie PCS was linked to the MFF, enabling all demographic and trapping information for an animal to be linked to the woylie PCS components for the same individuals (e.g. disease samples and field health checks).

3.2.2.3. Disease data collation

A template was provided to external collaborators for entering and returning results.

The aim being to;

1. standardise the format of the data;
2. make the process more efficient and reduce double handling of the data and;
3. most importantly to accurately match the sample results to the correct animal.

Cross-checking of samples was conducted with collaborators results and database records to account for all samples taken. This has required coordination with external collaborators. Attempts have also been made to cross-check samples that have not as yet been analysed.

Data validation of the disease data has primarily involved checking that the collaborators results are matched to the correct animal ID and that all results are captured within the database.

3.2.3. Results and Discussion

3.2.3.1. Aggregation of databases toward regional/corporate centralisation

The task involved close coordination and liaison between Peter Orell, David McKenzie, data custodians and WCRP team members to complete the aggregation process. The aggregation process was complex and difficult, requiring considerable time and effort for all those involved. The process took longer than expected due to the large number of problems with the data that were not envisaged.

The task was completed and MFF sent for meta-analysis on the 25/05/2007. The data sets aggregated into MFF are listed in Table 3.2.1. Almost 40,000 records were imported, and almost half of those were woylie records. This was a very good base for woylie meta-analysis, particularly when added to the approximately 5,000 woylie records already contained within Fauna File (PCS 990, Upper Warren 4,158).

In addition another 819 Karakamia, 49 Batalling, 32 Dryandra / Tutanning and 172 South Australian woylie records have been imported into the MFF from surveys conducted during the WCRP period.

A small percentage of records for each dataset could not be imported due to outstanding problems with the data records, in each case this was less than 5% of the data. For the Kingston dataset this was only 0.5% of the data. Data validation is still required by the custodians for the historical Boyicup and Yendicup datasets, as well as the Bushranger Yendicup and Yackelup data sets. Rating levels remain low for these data sets at present.

Importation issues

Ear tag clashes were the greatest problem in the importation process. The overlap in sites and survey periods between the different datasets caused problems determining whether ear tags were from the same or different animals. The occurrence of duplicated ear tag numbers was also a significant problem, particularly ear tags with no prefix. A central or at least co-ordinated control of the ear tags codes used by different operators is highly recommended to avoid this problem continuing in the future.

The dataset that was by far the most difficult to import was the Kingston dataset that was stored within a similar, but not identical database to Fauna File. The small differences created a number of issues for importation. For example, right and left ear tag order was not necessary in the

Kingston database. The animal was recognized either way. In Fauna File, however this was important in animal recognition.

There were also the inherent database issues of linked tables. The main table that holds the data is linked to several other background tables. If the person initially entering the data is not aware of this they can create erroneous records within these linked tables. These erroneous or "ghost" entries had to be fixed prior to importation, slowing down the process immensely.

The requirement to collect the original or 'parent' hardcopy data prior to importation of the actual trap data and restructuring of the data to meet the specific field criteria also slowed down the aggregation process.

Table 3.2.1. Datasets integrated with the original Fauna File datasets to create the 'Manjimup Fauna File' (MFF).

Data Source	Survey period	Custodian	Total No. records imported	No. woylie records
Kingston	1994-2005	Adrian Wayne	29151	>12000
Boyicup	1974-1986	Neil Burrows / Per Christensen	4088	3142
Yendicup	1977-1998	Neil Burrows / Per Christensen	1666	1397
Bushranger (Yendicup)	2000-2007	Bruce Ward / Graeme Liddelow	2165	940
Bushranger (Yackelup)	2000-2007	Bruce Ward / Graeme Liddelow	2055	590
Total			**39125**	**>18069**

NB: 1999 Boyicup data not aggregated into the MFF due to limited baseline information.

3.2.3.2. Creation of Woylie PCS database

Sections of the Woylie PCS database were operational in late 2006 and sections were added as designed by WCRP personnel. David McKenzie had completed all sections, including any required alterations by June 2007. All database sections were operational and running smoothly in mid 2007, in time for the bulk of the data entry to be completed and analyses to begin shortly thereafter. The development of Woylie PCS database based on current data requirements was comparatively easier and much quicker than the aggregation of data into MFF.

All components of the study, other than "demographics" were included within the Woylie PCS database (Figure 3.2.1). Table links with MFF enabled full capacity for the relational data to be integrated and interrogated accordingly. In eventuality the Woylie PCS database contains data from Perup (PCS and Upper Warren sites), Karakamia, Batalling, Dryandra, Tutanning and South Australia.

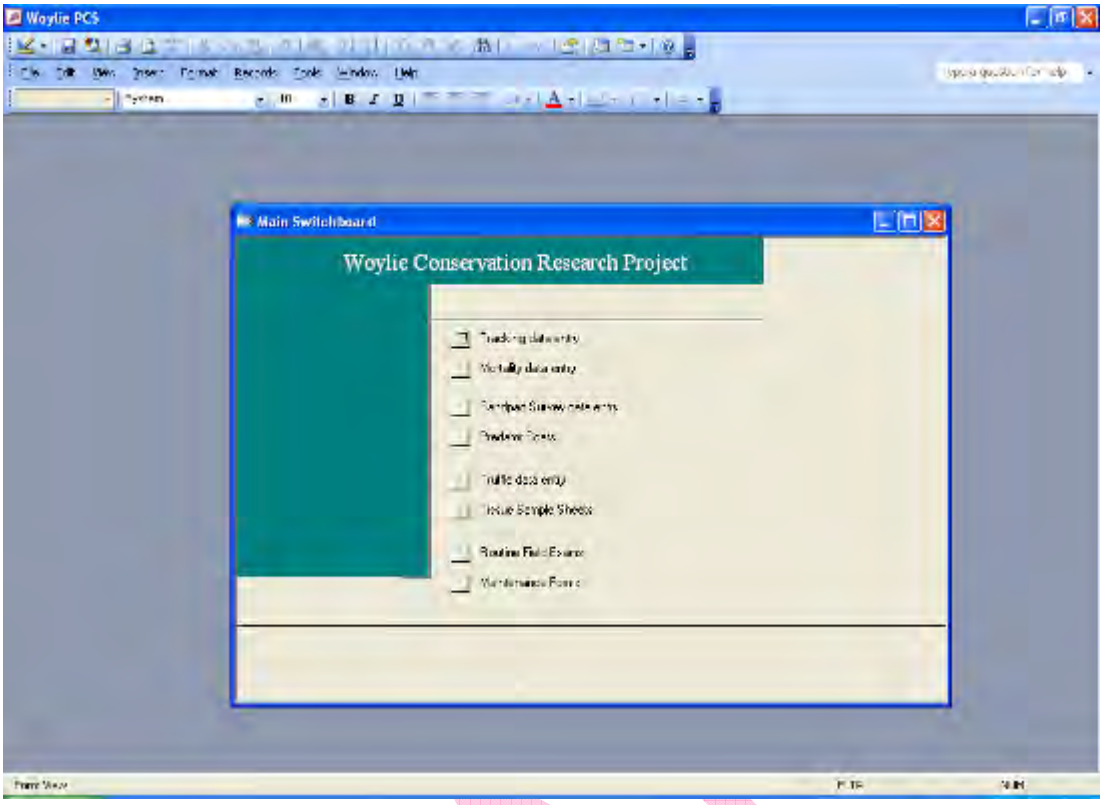

Figure 3.2.1. Woylie PCS database main menu.

3.2.3.3. Disease data collation

The results received from external collaborators and consultants have been entered into the Woylie PCS database by WCRP personnel and DEC casual employees. Therefore, for every animal that has had any sample taken there is a complete history of disease and demographic information.

The data has been cross-checked and validated with the custodians of the results as much as possible. Some work is still required before these datasets can be comprehensively interrogated with associated relational data.

Importation issues

The format of results from external collaborators and consultants were not consistently standardised as requested. It was particularly difficult to match samples to animals due to collaborators using their own reference numbers. The independent data was also returned in a variety of formats (i.e. Word documents, Excel spreadsheets, etc) and the structure varied as did the content. As a result, in the majority of cases the data had to be entered manually into the database rather than imported as intended. There are still current validation issues with outstanding records that have not as yet been positively matched to the animals from which they came. There will be an ongoing requirement for importation of disease results as they are made available from the external collaborators and consultants.

3.2.3.4. Data management and intellectual property

The intellectual property rights associated with the datasets provided by multiple custodians has been a high priority for the management of these datasets. This is especially important given the high level of involvement of individuals from a number of institutions and agencies. While fundamentally based on trust, co-operation and a conducive culture, the use of the data is managed through a data request process that registers the use, its purpose and gets approval from the original authors/custodians of the data (Volume 3 Data-use conventions and Data request form).

3.2.4. Future work

- To achieve the goal of centralisation of fauna data sets it is very important that unique ear tag prefix and numbers are used when conducting fauna surveys.
- Reinforcing and developing the standardisation of data collection methods and units of measure for fauna surveys.
- Database training or close supervision for personnel entering data to prevent the creation of "ghost" records within the database.
- Some improvements to the databases could be made to enhance the quality control provisions at the time of data entry and to facilitate data validation efforts.
- Ongoing maintenance of the MFF to both update and validate existing and historical data in conjunction with the custodians of the aggregated data sets.
- Web based Woylie PCS database that would allow external collaborators and consultants to enter their results directly into the database and enable them to cross check and validate their own data in relation to the individuals from which the samples were originally collected.

3.2.5. Conclusion

- Successful aggregation into the Manjimup Fauna File (MFF) of multiple, previously independent datasets associated with the Kingston Project, long-term monitoring of Perup, Bushrangers, *Western Shield*, DEC Fauna Management Course, and Donnelly District fauna activities.
- The Woylie PCS database was created to centralise all data pertinent to the WCRP including survival and mortality data from radio-collared woylies, predator, resource, and disease components of the study. This database was linked to the demographic data within the MFF to provide an easily obtainable complete history of disease and demographic information for each individual animal.

Practical advantages of the aggregation include;

- Comprehensive relational history of individual animals involved in multiple fauna programs.
- Substantial progression towards standardisation of data collection and management, including improved data quality control.
- Optimises the value of the data collected over more than 30 years.
- Facilitated efforts to diagnose the woylie declines in the Upper Warren, particularly by providing the data available for the subsequent meta-analyses.
- Provision of a useful model and case study for considerations of the development of a corporate-wide aggregation of similar datasets.

3.2.6. Acknowledgements

We would like to thank Sheryn Prior for her short but invaluable data management role and in particular her persistence and diligence with the preparation for importation of the historical Boyicup and Yendicup datasets. We would also like to thank Lauren Daubney and Verna Tunsell for their data entry contribution.

Marnie Swinburn, in her roles as Deputy Disease Coordinator and External Programs Liaison, was responsible for liaison with external collaborators and ensuring provision of data for both the Woylie PCS as well as to the requesting collaborators.