Project Plan CF 2015-024

BioSys – the Western Australian Biological Survey Database

Ecoinformatics

Project Core Team

Supervising Scientist	Paul Gioia
Data Custodian	Paul Gioia
Site Custodian	

Project status as of July 13, 2018, 10:58 a.m.

Update requested

Document endorsements and approvals as of July 13, 2018, 10:58 a.m.

Project Team	granted
Program Leader	granted
Directorate	granted
Biometrician	granted
Herbarium Curator	not required
Animal Ethics Committee	not required

BioSys – the Western Australian Biological Survey Database

Biodiversity and Conservation Science Program

Ecoinformatics

Departmental Service

Service 5: Conserving Habitats, Species and Ecological Communities

Project Staff

Role	Person	Time allocation (FTE)
Supervising Scientist	Paul Gioia	0.5
Research Scientist	Florian Mayer	0.2

Related Science Projects

CF 2011-106 Online GIS biodiversity mapping (NatureMap)

Proposed period of the project

June 30, 2015 – None

Relevance and Outcomes

Background

One of the department's corporate goals is to conserve biodiversity. A key strategy for achieving this is acquiring scientific knowledge to underpin decision-making. The department therefore invests heavily in ecological survey (from small area- inventory through to major regional surveys) and on-going monitoring projects from which data are collected through field observation and analysed to produce new information. Because of the size of Western Australia, and the inaccessibility of many sites, data collection is typically the most expensive component of producing new information and understanding, and the datasets are often irreplaceable.

However, the majority of datasets collected through survey are not lodged in a corporately accessible and managed environment. While the department maintains some corporate scientific and management databases, these capture only a small proportion of data generated through survey, probably less than 25%. The remainder of datasets are managed on a per-project basis, satisfying individual project requirements, but invariably without due attention to on-going data maintenance and availability. To compound the issue, an increasing number of scientists and project officers responsible for historical datasets are close to retirement age. The department therefore faces an increasing and significant risk of data loss.

Compounding the situation, data are stored in various ways ranging from spreadsheets to Microsoft Access, on desktops, laptops, or local file servers. As such, security and data protection are dependent on the data custodian, resulting in inconsistent practices and increased likelihood of data loss or corruption. In addition, data access is typically limited to local users, or restricted by vendor-specific applications. Departmental requirements for linking to related systems are generally not considered. And, finally, while survey methodologies are well understood, inadequate attention is paid to data collection and management standards (particularly ecological data, which still has developing standards), and therefore inconsistency and lack of documentation in how data attributes are defined.

Aims

- Build a central repository for storing, curating and distributing ecological data
- Minimise the risk of data loss to the department
- Increase access to ecological data

- Create a single point of truth for ecological data within the department
- Develop best-practice techniques for managing ecological data within a corporate context
- Facilitate the development of ecological data standards and survey protocols

Expected outcome

- Increased consistency in, and availability of, ecological data
- Data protected in perpetuity against staff turnover and change in storage technologies
- Improved security and backup for legacy and operational databases
- Improved integration with other departmental systems
- Improved capacity for reporting, research and analysis of observational data.
- Improved compliance with government requirements for knowledge management
- Improved credibility of the department in protecting and making data available
- Contribution to the public good by making all intellectual property publicly available, and encouraging co-development from sister organisations

In the longer term BioSys is expected to take over all or part of the role of NatureMap, built on modern, supported architecture.

Knowledge transfer

The novel approaches to ecological data management used within BioSys will have relevance to any biological survey database. BioSys is being developed as open source software, and intellectual property and source code will be made available to the general scientific and informatics community.

Interest has already been expressed by the NSW Office of Environment and Heritage in supporting and co-developing the application.

Tasks and Milestones

Develop minimum viable product (MVP)	30/06/17
Test database design against Pilbara and other key datasets	31/12/17
Implement interface to NatureMap for automated data consumption	30/06/18
Implement standardised ecological data model	31/12/18
Expose source code as open source and invite collaboration	01/07/19
Assess feasibility of offline data capture tool	31/12/19

References

Australian Venture Consultants (2012) Pathway to an Enhanced Western Australian Terrestrial Biodiversity Knowledge System, August, 2012, Perth, WA.

Salt, C., Burrows, N., Coates, D. & van Leeuwen, S. (2008) Vegetation information management system: the need for a new vegetation map of WA: Vegetation Mapping Workshop. Department of Environment and Conservation, 23-24 July 2008, Woodvale, WA.

Science Division (2008) A Strategic Plan for Biodiversity Conservation Research 2008-2017. Department of Environment and Conservation, Perth, WA.

Study design

Methodology

Design principles and approach for BioSys include two key attributes: *open source development*, and *scalability*. All intellectual property created during BioSys development will be made available through the open source

paradigm. While early stages of development will be kept in-house, once BioSys has reached a basic level of acceptance and maturity, the source code will be housed within a publicly accessible repository. Other developers and research institutions will be able to contribute to the code base so that the software will benefit from potentially many sources.

Key reasons for making the source publicly available include a) providing a public good benefit given that BioSys has been developed using public funds, and b) benefitting from the keen interest that other parties have already expressed, and the additional expertise they could bring to the project.

There has been a high level of difficulty in the conceptual design of the database thus far. Ecological data is, by its nature, complex. Additionally, survey protocols change over time as knowledge and techniques improve, resulting in different attributes being collected. And, finally, the lack of ecological data standards, together with a culture of independence amongst some researchers, contributes to even greater variation in survey protocols and data content and organisation.

This leads to a second key design attribute: scalability. It is a challenge to design a database that will scale to increasing data complexity and variability so that all ecological data can be captured and curated, while on the other hand providing reports and outputs in consistent formats for scientific research, conservation planning or environmental impact assessment, or consumption by other information systems. Traditional methods (e.g. relational databases) do not have the flexibility to accommodate complex ecological data. BioSys is therefore being designed using the most recent techniques for storing and updating semi-structured data that provide the required scalability and flexibility.

An inherent tradeoff with the above approach is that variability in data structure will be replicated within the database. Inconsistencies in how researchers manage their data will potentially be perpetuated. It is not possible to guarantee storage and curation of any dataset, while also guaranteeing those datasets can be consolidated into a single, monolithic, relational data model (e.g. for reporting or data exchange purposes). That is a much harder proposition.

The BioSys project is therefore a long-term project. In the first instance, risk of data loss must be mitigated. It is therefore more important that data is captured and protected within a corporate system “as is”, than attempting the harder task of data standardisation. BioSys will also allow data to be curated, the system potentially becoming the point of truth, rather than the desktop. Development will be incremental and adaptive - system capability and design will be enhanced as new datasets and survey protocols extend conceptual boundaries. In subsequent years, other capabilities will be added (subject to available resourcing), such as offline data capture tools that feed into BioSys.

Architecturally, BioSys is being developed as a web-based application using OIM-preferred and supported technologies.

BioSys is being implemented in a phased approach:

- Build database to support Kimberley Land Conservation Initiative (LCI) monitoring data and Kimberley Island survey data
- Release minimum viable product for acceptance testing
- Test and adapt database design by accommodating Pilbara and other key survey data sets
- Implement a standardised ecological data model based on outcomes from initiatives such as Essential Measures vegetation working group
- Evaluate feasibility of an offline data entry tool
- Invite co-development from interested parties

The phased approach described above is dependent on ongoing funding. Estimates below are based on current and historical funds provided from the Kimberley Land Conservation Initiative, at Director’s discretion, but this is not an indication such funds will continue to be provided in the future. There is the possibility projects dependent on BioSys architecture might subsidise aspects of BioSys functionality.

Biometrician’s Endorsement

granted

Data management

No. specimens

N/A

Herbarium Curator's Endorsement

not required

Animal Ethics Committee's Endorsement

not required

Data management

- Data will be managed using novel techniques to handle complex and highly variable data.
- All data will be stored within corporate infrastructure and benefit from standard disaster recovery processes.
- Data will be archived periodically and stored off-site.
- Data will be audited for changes at user-level

Budget

Consolidated Funds

Source	Year 1	Year 2	Year 3
FTE Scientist	0.5 FTE	0.5 FTE	0.5 FTE
FTE Technical	0.1 FTE	0.1 FTE	0.1 FTE
Equipment			
Vehicle			
Travel			
Other			
Total			

External Funds

Source	Year 1	Year 2	Year 3
Salaries, Wages, Overtime			
Overheads			
Equipment			
Vehicle			
Travel			
Other			
Total			