Bio data management in NZ: Current Practices and Options

This page is to give you an unbiased (as much as possible) overview of bio data management options in NZ. It is based on current practices in NZ.

This resource should help anybody considering an in-house bio data management system.

Feel free to add / edit!!!

Traditional occurrence biodata, especially in the freshwater and terrestrial space, focussed largely on determining species ranges. More recently, as the numbers of observations and interval they cover has increased, users are asking new questions of the data - where fitness for purpose can be problematic. Today, the key interest is in topics such as species occupancy and changes in range - eg: are endangered species ranges shrinking, are invasive species spreading?

So the "when" of an observation has become as important as the where. Also, absence information - determining where (and when) a species is known NOT to occur - is increasingly important to identift changes in species ranges. An absence observation needs more method metadata than a presence observation - as the user must determine whether or not the sampling methodology was appropriate for a particular species - so the absence is either unconfirmed or if the species could reasonably be expected to have been found if present. For example - nocternal species would not be expected to be observed during the day, or net mesh size may be inappropriate to retain small species, etc... Older observation data is often missing such metadata, and so is not suitable for repurposing in this way. New observation capture programmes should make sure to collect and provide such metadata to ensure they are fit for today's purposes. It is worth noting that in the marine space, many (but by no means all) species are still in the early phase of determining ranges, but this should not mean that less complete data should be collected).

In many cases, modern species observation programmes collect a wide range of data in addition to simple species observations. Mark/recapture programmes provide not just species presence data, but also provide individual "from here to there" data - often with no idea of what happened in between. Marine mammals are increasingly being identified as individuals from marks/scars, etc - so multiple and ongoing observations of actual individuals are collected - not just species, or just mark & recapture. Some animals (marine, freshwater and terrestrial) are being tagged with data logging tags, that either report via satellite, or can be found and data uploaded once they detach from the subject. These programmes collect substantial information as a timeseries while attached - so a more complex database design and application is required. Biodata are also captured as part of other research programmes (oil & gas seismic surveys for example) and increasingly, citizen science and NGO's are playing a part in capturing biodata (Naturewatch, eBird, Plant Conservation Network).

It is also worth noting that institutional databases are generally incomplete - and do not provide a genuinely national dataset. Biodata are often spread across institutions, and especially in the case of marine and bird species, they are international. Often a complete datasets can only be obtained from a global system (OBIS, GBIF, etc) which harvests all the institutional and national databases into a cohesive single repository. If you are implementing an institutional or national database, you should consider what level of interoperability is appropriate with other biodatabases, and how your data can best be provided to potential users, who may wish to combine then with observations from other sources for their purposes.

Genetic information is increasingly being collected, along with specimen data, so potential support for data sharing with the major international genetic database repositories may be required.

When choosing or designing a tool to manage biodata, all these issues should be considered - not just for the immediate needs, but also likely future needs. There are many available systems, some commercial products, some open source software and others which are developed in house to meet institutional needs. How well the software needs to comply with biodata standards such as IPT for Darwin Core Archives for submission to repositories such as GBIF, for example. Some tools are listed below:

- Commercial solutions for time series data management (which is not necessarily biodata observation data - but is related):
    - Hilltop?: well established time series management package; there has been some noise about adding bio data management component. http://www.hilltop.co.nz/
    - Kisters?: well established time series management package; there has been some noise about adding bio data management component. http://www.kisters.net/wiski.html
    - Aquarius?: well established time series management package; there has been some noise about adding bio data management component. http://aquaticinformatics.com/
    - EcoBase: established dedicated biodata management system, used by Auckland Council; http://www.iquest.co.nz/ecobase.php; contact Mike McMurtry <Mike.McMurtry@aucklandcouncil.govt.nz>
- IRIS: Who can add more info??
- CADDIS: MSACCESS based system to manage bio observations; now ported to SQL Server (Horizons/Cawthron)
- The "NIWA system" old "FBIS" /new "NEMO":
    - NIWA manages lots of primary bio observation data in legacy simple systems "excel/access" solutions. Given the diversity of our operations we think at this stage it is not feasible to implement a one.NIWA solution for bio data management.
    - Our implementation is therefore based on the premise that experts manage their data using "simple solutions" but we provide a "data warehouse approach" for aggregating / archiving the data into a central space.
    - To enable that, NIWA has developed an in-house postgis based data archive for all our sample data (note: just data archive / no management system!).

        - This is basically a database with some low-level web services connectors for data ingest and data query (through OGC services)
        - More information: http://www.niwa.co.nz/environmental-information/update/environmental-information-update-3-august-2012/integrated-and-easily-accessible-data-on-new https://teamwork.niwa.co.nz/display/NZEIIF/The+NEMO+Data+Model+and+XML+Representation
    - NIWA is currently working on providing documentation on the data model.
    - The data model and code are available without restrictions for anybody to adopt, however it is likely that extensive in-house (database and application) development will be required for redeployment.
- iNaturalist : free, open source citizen biodata observation software as used by Naturewatch
- AOL - Atlas of Living Australia - CSIRO developed as an Open Source biodata (species observation) portal - has been trialled by Te Papa for an NZ facility