Practical recommendations for improving institutional bio-data management and delivery infrastructure

The Bio-data Services Stack (BSS) project has worked with several New Zealand institutions to evaluate current bio-data management systems for Freshwater Fish, Freshwater Invertebrates, Pest Plants, and Birds observations and to plan how federation of this data can be achieved (with other providers in New Zealand). Here, we summarize some of the key findings and recommendations on how institutional bio-data management could be improved. The particular 'angle' taken in this evaluation is to look for consistency and robustness in the data management systems, to ensure data can be easily re-used in the future.

'Typical' observed weaknesses in bio-data management include the following.

- Metadata management to ensure information about the datasets (including the data collection protocols and dataset field name ontology), is available and regularly reviewed, is missing;
- Documentation of the data management systems and data models used, is missing;
- Daylight saving timestamp often used (and often not documented). While this might be not too critical for biological observations such as vegetative extent, it is highly critical for daylight-dependant observations such as bird song;
- Species names are weakly managed (e.g. not regularly validated and updated ad-hoc);
- There is no consistent site naming conventions and record identifiers;
- Some data are managed in excel with no clearly mandated and documented format (the format / data model is maintained by 'some person' in his head for instance); and
- No clear business path or agenda to modernise data management systems (for biological data).

This means some of the details of the datasets might get lost and / or rely on individual knowledge (risk of loss).

The following recommendations are made for standards/practices to be applied by each organisation.

1. Establish **an organisational dataset catalogue(s)**, for maintaining core metadata about data assets in a consistent way (and compliant with national and international standards) across the organisation. This would simplify (automate) the process of publishing (bio) data assets with the attached metadata.
   (Alternatively, for the specific purposes of bio-data only and the BSS project, an IPT server can be used to manually enter and manage the dataset metadata. Part of that would also be the assignment of unique (metadata) identifiers to the datasets.)
2. **Establish a consistent, organisational site naming convention** across various data collection systems. This would simplify and automatize combination of data from various datasets (for example water quality with invertebrate and fish observations).
3. Establish a consistent organisational way / organisational convention and policy to create **unique record identifiers** for observational data across organisational systems (for environmental observations). This would enable easy 'whole agency' traceability of observational records.
4. Establish an **internal policy / best practice for consistent time stamping** of observational data. Often systems use daylight saving time (partly without being recorded in the metadata) which should be discouraged and removed / converted in the data. It should be mandated that always NZST is recorded and the recorded timestamp includes the UTC offset and is completely described in ISO8601 format in any data management system.
5. Establish **best practice for consistent geolocation referencing** in place. From our observations most of the systems are using New Zealand Transverse Mercator (NZTM) as the geolocation reference system. A practice of using one georeference system within an organisation should be continued and enforced as an institutional policy. Additionally the EPSG code should be recorded in each data management system (for completeness). We note that a local reference system like NZTM might not appropriate because it does not cover offshore Islands. Use of a global system (like WGS 84) might be more appropriate.
6. Establish one institutional **central taxa name management system**[1] (with connections to a reference name service, as established by the New Zealand Organisms Register NZOR), which is connected to the different observational databases. This would include a management role for ensuring data integrity. This would centralize the current practice for each system to maintain its own 'taxa lists' and would require an organisational change.
7. **Assess all current 'manual' / desktop based data management systems based on spreadsheet and other such like software (e.g. MsAccess).** The practice of using excel spreadsheets means high risk to data loss (physical and/or semantically). For some data management systems (managed by one person) it might be appropriate to use such desktop software, however, in this case, very good documentation of the used data model / format, and vocabularies, etc. is required ('Nothing / no assumption shall be stored in heads').
8. Establish **formal (and well documented) data quality assurance processes**. This especially applies to Weeds data where no formal data control system is in place.
9. Improvement of **documentation of data management systems** by ensuring best practice and templates for documenting data streams and archival procedures are in place, maintained, and followed. This links with recommendations (1) and (7). (This would mitigate risk of data management systems knows only to a few individuals.)
10. **Establish a formal Data Publishing Policy** which specifies the rules the Organisation sets around to whom and under which conditions data is released and published on the Web.

As one can see, these recommendations are mainly centred on establishing new or 'tighter' policies, procedures, and practices in data management. While it is acknowledged that this comes at a cost, we believe that in the long term the organisation will gain from that through much improved data security and easier data exchange.

Generally the workshops carried out through the BSS project demonstrated the usefulness of bringing together different data managers (even within and organisation) in a workshop environment to discuss their data management issues and give people time to ‘think outside the box (of day-to-day routine)’ for the overall benefit of the company. This is something to be encouraged.

It should be noted that implementing (some of) above recommendations is dependent on the specific situation within an institution and planning and costing would require further scoping work. These recommendations shall encourage institutions to invest in such an exercise and provide a starting point. The Bio-data Management Guide[2] developed as part of a separate TFBIS project is another valuable resource for such an exercise.

---

[1] As a general comment, the institutional use and maintenance of controlled and preferentially ‘published’ or authoritative vocabularies, look up lists, of which species names are a special case is highly important (and often overlooked).
[2] http://dataversity.org.nz/guide/