The New Zealand Bio-data Services Stack

Infrastructure and experiences from stakeholders

Final Report of Phase Two of the New Zealand Bio-data Services Stack Project

TFBIS project 299: “A national network for connecting and mobilising primary biodiversity data”

https://teamwork.niwa.co.nz/display/NZBSS

June 2015

Jochen Schmidt, Nick Spencer, James Lambie, Mike McMurtry, Alistair Ritchie

Contents

New Zealand Bio-data Services Stack (BSS) – Facilitating biological data exchange across New Zealand

During 2014 the TFBIS funded "New Zealand Bio-data Servcies Stack" project (TFBIS project 299: "A national network for connecting and mobilising primary biodiversity data") has worked with key stakeholders on defining what a national Bio-Data Infrastructure (BDI) for better integration of New Zealand biological data derived from multiple distributed sources should look like. The BSS project restricted its scope on observed species abundance data, and focused on four 'species domains': birds, pest plants, freshwater fish, and freshwater invertebrates. These were selected as they were perceived by the New Zealand bio-data community (as represented at the 2013 Dataversity workshop) as most important for the countries environmental management needs. Please refer to the project WIKI (https://teamwork.niwa.co.nz/display/NZBSS) for background on the project.

The BSS project aims to demonstrate inter-agency bio-data federation for selected data sources. The BSS project consists of three phases. In Phase One we identified the BSS requirements focussed on the identified case studies and related data sources. As part of that, data sources and barriers were identified, and some initial draft national standards and guidelines for data exchange developed. The report on BSS Phase One has been published as Lambie et al. (2014) and is available on the project WIKI. In Phase Two we have been further developing and applying relevant infrastructure components, and in particular have been working with data providers through a number of workshops on enabling them to publish data to standards. We have demonstrated a working prototype of the Bio-data Services Stack through provided services and client tools. Phase Three is dedicated to the refinement of the developed standards and broader uptake of BSS with partners.

This report documents Phase Two of the project, which is aimed at defining the national Bio-Data Infrastructure (BDI) required to meet the BSS needs (as identified in BSS Project Phase One), and demonstrate the feasibility of the BDI through a working prototype of services and clients. During this Project Phase we have worked with a number of key agencies in federating some of their data according to BSS guidelines. Here

- we report on the required structure of a BDI to support a BSS;
- we demonstrate the working prototype of occurrence bio-data federation; and
- we document some of the learnings we have made during our work with some key stakeholders.

BSS Project Phase Two Work Programme

As a result of BSS Project Phase One, a 'strawman' of a BSS standard was developed (can be found on the BSS WIKI, https://teamwork.niwa.co.nz/display/NZBSS). As part of BSS Project Phase Two, this strawman was evaluated and further improved through stakeholder feedback and a series of workshops (dates and venues can be found on the BSS WIKI, https://teamwork.niwa.co.nz/display/NZBSS). During these workshops we worked with a series of stakeholders on what needs to be done to federate their data sources according to our developed code of practice and guidelines. The experiences provided invaluable information which helped refining the BSS Code of Practice (see next chapter), and gave feedback to stakeholders in the form of general learnings and specific feedback for individual stakeholders.

Key goal of the work programme was to achieve the following outputs.

- To document a 'BSS Draft Code of Practice' that is required to support occurrence bio-data federation. This Code of Practice (i) documents the standard formats for data federation and (ii) provides guidelines for data providers and data consumers on how to implement these standard formats into their workflows.
- To demonstrate that the developed BSS Code of Practise is suitable for helping data providers publishing their data sources. We did this by working with select project partners providers through what needs to be done to change internal institutional workflows to enable an agency and to document the learnings from that (for the project implementation phase)
- To demonstrate that the developed BSS Code of Practise is suitable for enabling data consumers to aggregate occurrence bio-data sources from many data providers.
- To document the wider components that could potentially form a national Bio-Data Infrastructure (BDI). We note that occurrence bio-data is very restricted in its nature, and any next stages of work would need to resolve how to manage much more complex sets of bio-data (which the current project has not focused on).

The BSS Draft Code of Practice

A 'BSS Draft Code of Practice' was developed which

- defines how bio-data shall be published as part of the BSS Infrastructure work stream / BSS Project Phase Two; and
- provides practical guidelines for agencies how to publish their bio-data.

For practical and efficiency reasons we based this version on an existing international biodiversity data standard (Darwin Core - DwC) and an existing data publishing service framework (Integrated Publishing Toolkit - IPT). IPT was developed by the Global Biodiversity Information Facility (GBIF) as their key mechanism for harvesting biodiversity data from around the world. IPT uses DwC, developed by the Taxonomic Database Working Group (TDWG) as its primary data format for exchanging biodiversity data from provider to GBIF. We worked on adapting these to BSS requirements by developing the BSS Draft Code of Practice as a refinement using these standards and tools. This included linking relevant taxonomic names fields within the DwC format to the New Zealand Organisms Register (NZOR) to ensure consistency in the described observation methodology.

The BSS Draft Code of Practice is available as a 'living document' under

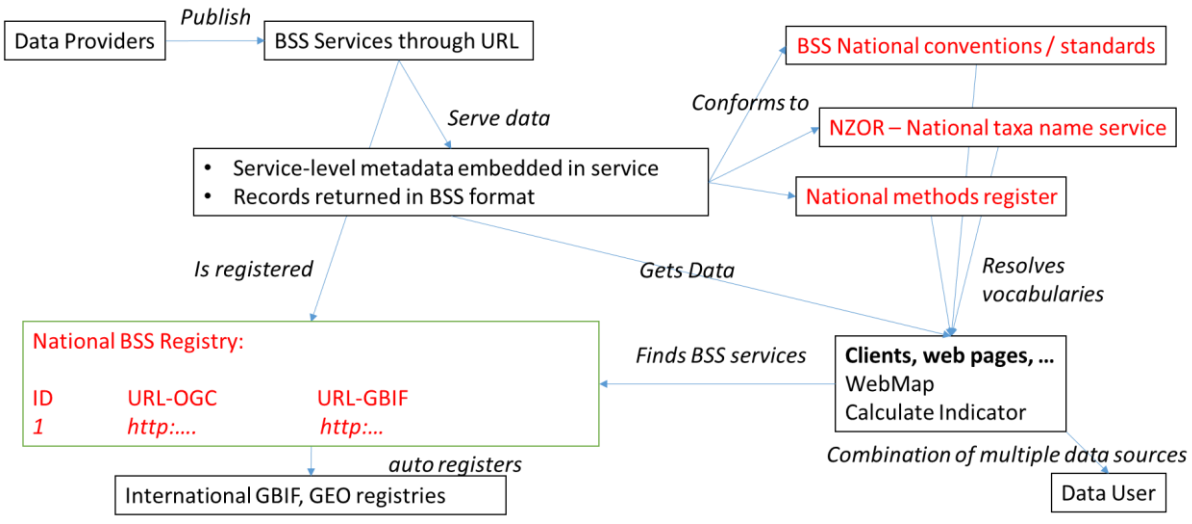
https://teamwork.niwa.co.nz/display/NZBSS/BSS+Infrastructure%3A+BSS+Standard+and+Data+Infrastructure+Components

and

https://docs.google.com/document/d/1Z_sUN51YImh9vdPDew9_dxBMpmizCb_A_oWANSvkdXw/edit?usp=sharing

The process of the development of the 'BSS Draft Code of Practice' is described on the BSS project WIKI under:

https://teamwork.niwa.co.nz/display/NZBSS/BSS+Infrastructure+Work+Stream.



Graphical illustration on how the BSS system works (BSS products are in red).

BSS Draft Code of Practice Critique

As part of the BSS Project Phase Two we critically reviewed the BSS Draft Code of Practice throughout the process and identified issues including the following.

- The DwC data standards are suitable for certain occurrence bio-data, but may have limitation that need to be further tested when it comes to more complex (e.g. hierarchical) bio-data types, such as Tier 1 monitoring.
- While good for transporting simple primary abundance data, measured in species count, many data providers wanted to transport more data, often specific to a particular type of observation, for example observed specimen length, weight and the such as well as environmental conditions observed during sampling, e.g. weather, habitat, flow conditions. While the Darwin Core "Measurements and Facts" extension offers a solution for that, it does not fit well within a standards based framework that supports ontologies.
- The DwC standard works well for the distribution and sharing of occurrence bio-data and allows large volumes of data to be processed by automated methods. It also usefully presents this information in a human readable form. However, DwC and does not inherently have the necessary structure and capability to provide true machine-to-machine reasoning and data transfer. Certain properties are ambiguously defined, with vague content models. Others rely too much on free text, rather than controlled content (see previous point). This can introduce uncertainty. Domain restrictions can mitigate some of these impacts.

Based on these findings we will operate a dual approach. Firstly, we will continue to support and test the applicability of DwC for a wider range of bio-data uses case as we continue with workshops and end-user engagement. Second, and in parallel, we will investigate how a more generic observations data standard might the address some of these questions. We will compare and contrast the results of this investigation and use this analysis to inform our final recommendations.

Implementation with Stakeholders

As part of BSS project Phase Two we conducted a number of workshops with key stakeholders for enabling them to publish their data using the BSS Draft Code of Practice. Dates and venues are recorded on the BSS project WIKI.

As part of these workshops

- we brought IT experts, database specialists, and scientists from organisations together;
- we reviewed the organisational data collection and management infrastructures; and
- we mapped the held information onto the formats defined in the BSS Draft Code of Practice.

As output of the workshops

- we provided specific recommendations for the relevant organisation on how they can improve their data management infrastructures (see next section); and
- we provided specific instructions on how the data provider can publish their data to match the requirements of the BSS Draft Code of Practice.

Generally these workshops were welcomed by the agencies not only for enabling them to work towards standards based data federation, but also as an opportunity to take a 'hard look' at their internal data infrastructures with experts from many fields coming together for this exercise. During 'normal business' this rarely happens.

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

[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).

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.

[2] http://dataversity.org.nz/guide/

BSS Prototype Implementations

As part of the BSS Project Phase Two we have implemented and tested a number of BSS compliant services as well as clients.

- Horizons Regional Council has setup an IPT prototype server.
- Landcare Research has setup an IPT server for their terrestrial data.
- NIWA has setup and IPT server for their freshwater water data.
- The BSS team is working with Auckland Council and Environmental Southland on publishing their data through the NIWA and Landcare's IPT servers

In addition to those services the following work was done.

- NIWA supported the implementation of a demonstration QGIS client which enables interested parties to connect to IPT servers and display / analyse data from many sources in a geospatial GIS framework.
- The BSS project worked with the AgResearch - mediated 'National Weeds Database' project[3] on evaluating the developed BSS standards and services.
- The BSS project worked with the LAWA project (Horizons technical personnel is part of the LAWA and BSS project team) to ensure the BSS solutions are also applicable for the requirements of the LAWA project.
- NIWA also has worked over the last year on publishing its (and some of MPIs) marine bio-data through an IPT server and thereby demonstrates feasibility of re-use the information infrastructures developed as part of this project for marine data.

[3] Project Leader is AgResearch's Graeme Bourdot

A National Bio-Data Infrastructure

While the BSS project has the goal during its limited lifetime to ***demonstrate*** federation of occurrence bio-data within a limited scope, generally our goal in New Zealand is to work towards a persistent, sustainable, and well-supported and adopted national Bio-Data Infrastructure (BDI). A BDI should include required national / central infrastructure components, standards maintenance work and oversees adoption of data federation from a national perspective (based on the model of the national Spatial Data Infrastructure SDI as lead by the New Zealand Geospatial Office NZGO). It should also be capable to serving a wide range of bio-data types rather than just occurrence data *per se*.

Recommendations for implementation of a National Bio-Data Infrastructure based on the BSS work so far include the following.

It is recommended to setup a national infrastructure which

1. supports the further development and ongoing maintenance of bio-data standards (as demonstrated by the BSS project), this ensures that data providers publish their data to standards;
2. establishes and maintain a national registry service for bio-data sources / service endpoints[4], this ensures data sources can be discovered by end users;
3. supports a national facility for maintaining taxonomic concepts, as implemented by the New Zealand Organisms Register (NZOR), this enables organisations to publish their species data to a common standard;
4. establishes and maintains a national vocabulary / register for bio-data methods[5], this enable organisations to annotate their bio-data with a common methodology;
5. establishes and maintains a national vocabulary for observation names and units, this enables organisations to publish their observations to a common standard; and
6. establishes and maintains a national framework including a linked data URI policy to support national vocabulary services[6].

We note that

- This infrastructure needs to include clear mandates (even legislation?); clear responsibilities; and resources.
- The infrastructure needs to include supporting training programmes on data and standards skills.
- The infrastructure needs to include supporting communications and audit programmes to encourage uptake and give feedback to contributors on compliance.
- Many of the mentioned components functions / components are generic in nature (in particular (2) and (6)) and can be re-used in the context of a wider "Environmental Data Infrastructure

[4] A draft paper on principles for setting up a national "Linked Data Registry" services has been prepared by the BSS project: https://confluence.landcareresearch.co.nz/display/SBINTEROP/Linked+Data+Registries

[5] A demonstration vocabulary service has been setup as part of the BSS project under: http://test.data.scinfo.org.nz/x/def/bss/protocol/invertebrate/C1-P2-QC2

[6] A draft paper on principles for setting up a national "Linked Data URI Policy" has been prepared by the BSS project: https://confluence.landcareresearch.co.nz/display/SBINTEROP/Linked+Data+URI+Policy

EDI", and combined with the SDI work that is going on through NZGO. The BSS project has started conversations and hosted a workshop with NZGO on that matter.

- Any data infrastructure should strongly support machine-to-machine interoperability to ensure automatic processes can be developed based on it.

It is worth noting that this is the function of the national GBIF node in most countries that are signatory to GBIF. NZ is a signatory but without existence of such a node. Hence any BDI development in New Zealand should be explicitly include, as part of the relevant governance infrastructure, setting up an official GBIF node to support New Zealand's international commitments.

BSS Project Phase Two: Conclusions

The work carried out during the BSS Project Phase Two “Infrastructure” was invaluable for a number of reasons.

1. It provided the partnering agencies opportunities to have a ‘hard look’ at their bio information management practices across a range of often segregated institutional data infrastructures. Many workshop participants highlighted the usefulness of the BSS infrastructure workshops to find out ‘what others do in their organisation’ that deal with similar data. This highlights the issue, that often data owners in agencies are engrained in their individual day-to-day operational activities that they do not have time and resources available to review their data management practices or look at data standards. Hence we recommend that agencies with an interest in improved data management, integration and federation should make time for relevant staff (including subject matter experts and Information Technology experts) available for regularly reviewing and improving their data management practices. This could be facilitated by (a range of) national training programmes on ‘data (standards) skills’.
2. We provided specific feedback to partner organisation on how their bio-data management could be improved to enable them to be part of a federated data infrastructure.
3. It lead to an improved and joint understanding of what is required to provide a national Bio-Data Infrastructure (BDI) to support federation of bio-data.

BSS Project Phase Three: Recommendations

Based on the learnings from the first two BSS project phases, the following work programme for BSS Project Phase Three is recommended.

1. Scope a BSS profile for bio-data federation based on OGC and linked data technologies for a wider set of data types and use cases.
   During Project Phase Two we decided on using GBIF compliant standards and systems for our BSS demonstration work.
   Developing a BSS profile based on OGC and linked data technologies will
   a. enable closer integration with geospatial information systems (GIS) widely used for data management;
   b. enable direct integration of bio-data with other environmental data types;
   c. enable closer integration with a national Geospatial Data Infrastructure (SDI) as developed by the New Zealand Geospatial Office (NZGO);
   d. enable more complex bio-data types and use cases; and
   e. enable the use of linked data approaches that might complement OGC developments.
2. Continue to work with the project partners on data federation.
   a. Hold at least six more workshops with data providers on enabling their data federation;
   b. Work with AgResearch on setting up a demonstration weeds consumer service;
   c. Fully develop a QGIS IPT client; and
   d. Work with LAWA on setting up an operational consumer service.
3. Develop a business case and strategy for implementing a national Bio-Data Infrastructure (BDI).