The Biodata Services Stack

Context, Case Studies, and Strawperson

Report of Phase One of the Biodata Services Stack Project

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

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

December 2014

Authors:

James Lambie (Horizons Regional Council), Mike McMurtry (Auckland Council), Jerry Cooper (Landcare Research), and Jochen Schmidt (NIWA)

Reviewer: Jane Robbins (NIWA)

Contents

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

Biological diversity conservation and sustainable resource managers require better integration of biological observation data derived from multiple, distributed sources to provide the necessary complete national evidence-base to enable effective and comprehensive decision-making. Biodata are currently created and stored in various organisations, in a wide variety of forms, using a variety of data standards, and with varying – often undefined – accessibility.

The BSS project will demonstrate inter-agency biodata federation. 'Data federation' is defined here as the situation where the data is collected and archived in various organisations according to agreed standards and made available by the organisations through agreed standard web services over the internet; and various client services can (automatically) 'pick' the data up ('consume the data') from many organisations for a multitude of uses. A number of elements are fundamental for a functioning national data federation system including the following.

1. The development, implementation, and maintenance of national data exchange standards.
2. The development, implementation, and maintenance of a national data infrastructure (such as national registries, vocabulary services, etc.).
3. The development, implementation, and maintenance of national data collection and archiving standards for the relevant data collection systems (to ensure the federated data can actually be meaningfully joined together).
4. National governance, rules, and mandates around the institutional adoption of any standards and best practices.

As part of the BSS project, we will address the following issues.

- We will identify and evaluate use cases, data sources, and barriers, and develop draft codes of practice and guidelines for **data exchange (the first item as per above).**
- We will identify required national **data infrastructure components (the second item as per above)**.
- We will demonstrate through a set of proof-of concept services (setup with our project partners) that a national infrastructure for seamless data exchange can be build.

It should be noted that as part of the BSS project we will not specially evaluate and address issues related to nationally incomplete and inconsistent data collection systems (third item as per above); we also cannot address the issue on how to setup a national data infrastructure (item four as per above). However, the project outputs will provide important information for those items.

The BSS project consists of three phases. In Phase One we identify the BSS requirements focussed on a limited set of identified case studies. In Phase Two we will further develop the required infrastructure and demonstrate a working prototype of the Biodata Services Stack. Phase Three is dedicated to the broader uptake of BSS with project partners. The overall outcome of the BSS project is to prove that a New Zealand federated biodata infrastructure can be built and through the BSS demonstration services get national momentum towards implementation and ongoing support for such a data infrastructure.

The BSS project is focused on identifying commonality of data collection and use between local government agencies (Regional Councils and Unitary Authorities), the Department of Conservation, NIWA, and Landcare Research. It is also limited to four biotic domains – birds, weeds, fish, and aquatic macro invertebrates. The primary output from the BSS project is the definition of what is required to enable consistent federation of 'simple' biological abundance observations - what was observed, how much, how, where, by whom, and when. As a by-product, the project delivers proto-type federated data-stores or information services to support subsequent decision making. We have limited the project to this scope to enable us to get tangible results with the limited project budget and timeframe. This scope for the BSS project was defined in a consultative process as part of the Dataversity Workshop 2013, based on national priorities on which the workshop participants agreed (http://dataversity.org.nz/2013workshop).

Within the scope of the BSS project, and the focus on specified case studies, we will investigate the federated infrastructure necessary to support the mobilisation and integration of simple presence, absence, and count data. One of the functions of the BSS project is to enable data federation by ensuring the method of measurement and sampling are specified in association with the data, and to promote standards for data (field) content. It is not the function of the BSS project to develop, endorse, or promote particular methodologies for measuring, monitoring and reporting biodiversity data. However, in order to limit the complexity of the project, the project will be confined to sharing data that conforms to a limited number of specified methodologies. The degree to which the project can facilitate exchange of sample-dependant complexity (such as a BSS that is able to handle nested plot observations or handle presence represented in GIS polygon format) will be limited.

Vision: by 2016 New Zealand has developed an operational "**Biodata Services Stack (BSS)**" supporting a specified range of biodata collection methodologies.

The term 'stack' refers to a set of system components, which are applied consistently (see Figure 1).

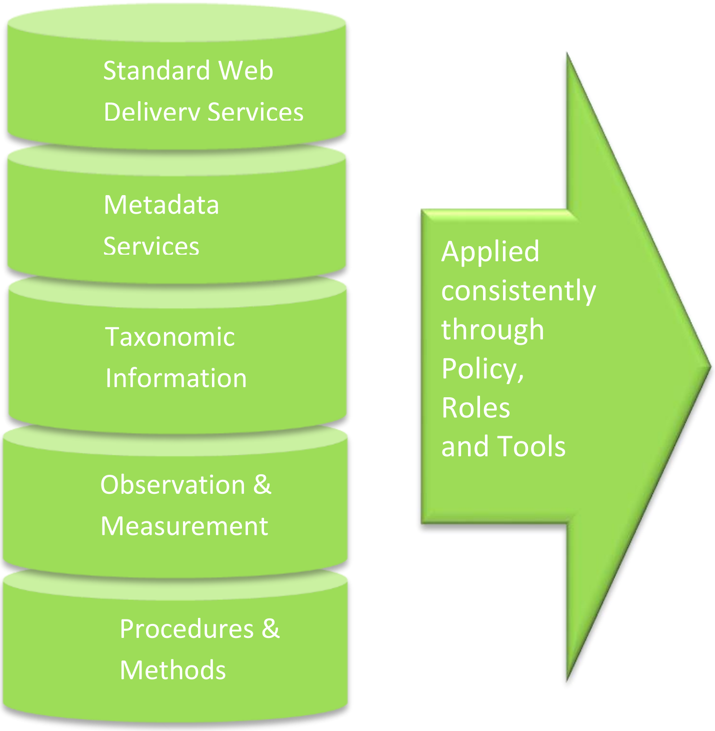

Standard Web Deliverv Services
Metadata Services
Taxonomic Information
Observation & Measurement
Procedures & Methods
Applied consistently through Policy, Roles and Tools

Figure 1: The BSS 'stack' of services, and their application.

Such a stack, if nationally implemented, ensures that, across NZ:

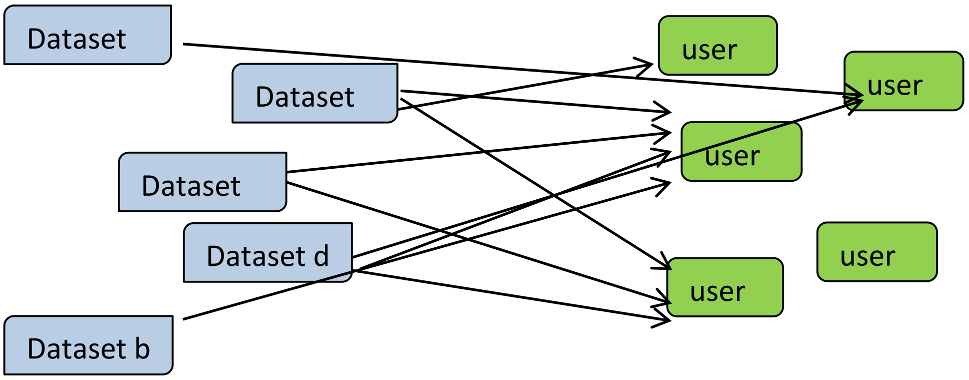
- **guidelines and procedures** are agreed and nationally maintained to ensure that (key elements of) primary biodata can be archived and published consistently, in any organisation.
- **support is guaranteed for national systems** (e.g. vocabularies, standards) for archiving and publishing of primary biodata conforming to specified methodologies as part of a national biodata infrastructure.
- **published biodata is discoverable and accessible** via standards-based web-services to enable data sharing and seamless (machine-to-machine) mobilisation.

The BSS project progresses the **New Zealand Terrestrial and Freshwater Biodiversity Information System's (TFBIS)** goal of a national network of connected and mobilised primary biodiversity data.

BSS facilitates the connection and mobilisation of biodata by providing coordination between data creators, data managers, data providers and data users.

Without BSS, datasets are likely to be generated according to ad-hoc formats, and lack standardised structure or format (Figure 2a).

With BSS, a set of data associated with a defined methodology is captured, together with appropriate metadata, and with content conforming to agreed data standards and vocabularies, and exchanged using agreed communication protocols. The BSS concept is validated via a number of case studies (Figure 2b).



Data Sources:

- different formats
- different owners
- different organisations

Data Compilation:

- ad-hoc
- manual
- repeated
- costly!!

Figure 2a: 'Life before BSS': datasets are inherently variable, and the transfer of data to data users has little or no standardised or consistent structure

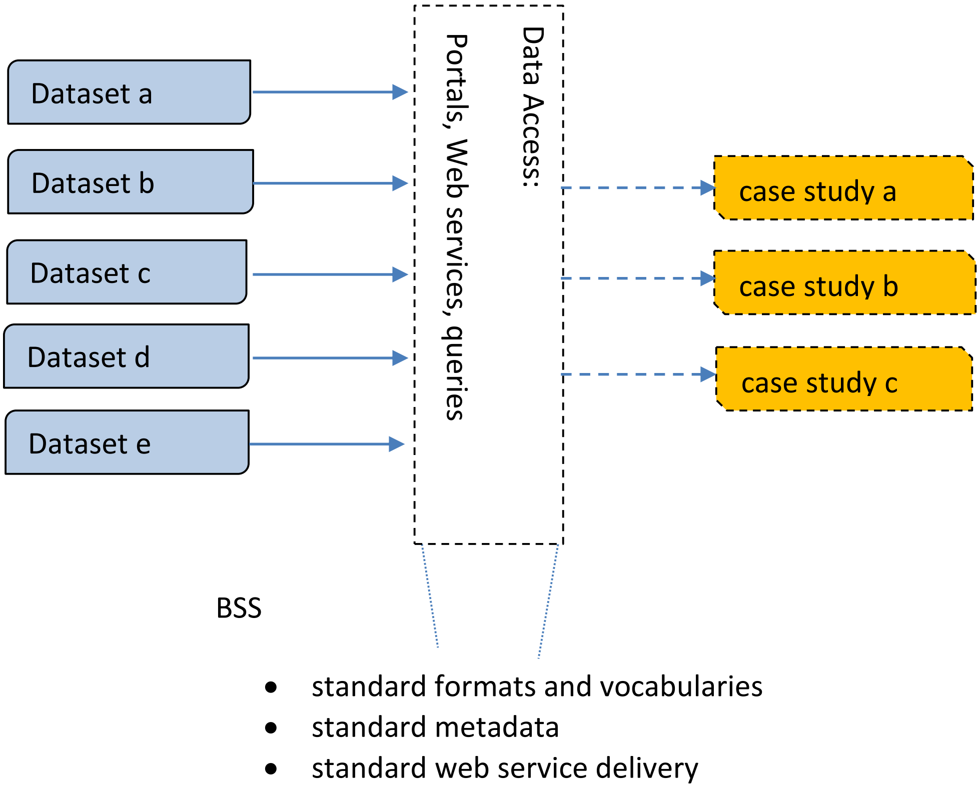


Figure 2b: ‘Life with BSS’: dataset contents are understood, enabling access to (and mobilisation of) data.

We would like to emphasise that the BSS project is not dealing with issues related to incomplete or inconsistent national data collection; it is restricted to a number of specific case studies and related data domains (as discussed); and it is a demonstration project with a finite lifetime. Further, this specific BSS demonstration implementation is not a generalised framework for mobilising and integrating all potential types of biodiversity data, hence only a first step towards a generic ‘biodata infrastructure’. While some of the components included in the BSS demonstration project, such as national method vocabularies services, national registry services, and national taxonomic services, apply also to more generic biodata, such a framework would require additional investment, research, testing, implementation and maintenance. For example, some key elements of a more generalised framework would include the development of a data exchange protocol for more complex site-based species co-occurrence data, such as that stored within the NVS databank, and the ability to share and integrate DOC Tier 1 survey data. In addition, BSS is intended to facilitate data access, standardisation and integration but does not specify how such data should be used. Enhanced data-exchange and national analytical services are just two examples of future infrastructure which only becomes possible once a BSS is fully developed and deployed.

The BSS Project: Phase One

The first phase of the BSS project is to identify and analyse use cases suitable for demonstrating data federation and exchange between agencies. This project phase defines the requirements for a Biodata Services Stack. The outputs are:

1. Clearly identify and document use cases demonstrating the mobilisation of species distribution data and barriers to doing that, in four biotic domains – birds, weeds, fish, and aquatic macro invertebrates; and
2. Scope the required solutions and BSS infrastructure components (data standards, guidelines for assembling data to be exchanged, vocabularies for standardising data content, etc.).

This report delivers two items for each domain: specific **Context** (defined need for federation given the current landscape of data collection and use across participating agencies), and a **Case Study** that demonstrates the value and need for a BSS (Output 1). The report also scopes a draft (a '**Strawperson**') of minimum data deliverables applicable to all **Case Studies** (Output 2).

The research methodology to identify the **Context** and **Case Studies** involved the following steps.

- Initial scoping of the BSS project and potential case studies was carried out through the Dataversity workshop in October 2013 with a range of stakeholders (stakeholders included DOC and local government agencies, as well as other central government agencies, NGOs, and private sector interests).
- Detailed biota-specific questionnaires were distributed to key DOC and local government contacts to test the validity of the Context, explore the value of proposed Case Studies, identify sources of data, ascertain current data management behaviour, and identify key potential barriers.
- Six regional workshops with local government stakeholders (see BSS WIKI for dates and locations, https://teamwork.niwa.co.nz/display/NZBSS) were held to re-validate the main findings of the questionnaires in terms of the Context and most useful Case Study for each biotic domain, to further identify sources of data, and to develop draft data standards for agreement/endorsement.
- A final validation (by DOC and other central government stakeholders) of the Context, chosen Case Studies, and draft BSS key data standards was carried out through a workshop in July 2014 (see BSS WIKI, https://teamwork.niwa.co.nz/display/NZBSS).
- During the BSS project, the relevant stakeholders were kept informed and feedback and review was sought through the BSS WIKI (https://teamwork.niwa.co.nz/display/NZBSS), regular emails posts and updates through the Dataversity Forum.

Contexts and Case Studies

This section captures the essential elements[1] of the **Context** and identifies the BSS **Case Study** for each of the biotic domains. Further detail on the Case Study and its product, the things BSS will help deliver on, alternative Case Studies proposed, and the data available can be found in the Appendices specific to each domain. We would like to emphasise again that the evaluations done as part of the BSS project focus on the data federation of observed species occurrence data.

[1] A draft Context for each domain was posed with the questionnaires. For brevity, the Context is paraphrased in bullet form. The Contexts have also been enhanced as a consequence of workshop discussion and further thinking.

The Domain of Birds

Case Study

Can access to dynamic data on bird distributions support the generation of trustworthy indicators for sustainability assessment and reporting?[2]

Current Context

- There are national schemes for specific birds or specific case studies, but no national standard protocol for monitoring common native birds is applied to all situations.
- There are a plethora of New Zealand databases and institutions handling bird observation data using a number of different sampling and observation protocols.
- DOC hosts a project for establishing the five minute bird count (5MBC) methodology (as well as other methods) as a common (standard) protocol, supported by a specific database.
- DOC uses the 5MBC method for DOC Tier 1 biodiversity monitoring and reporting.
- Many regional councils collect bird data already to support regional requirements for biodiversity assessment. There will be peer expectation for other councils to do so under the proposed regional Tier 1 biodiversity indicator for avian representation.
- Commonality of sampling protocols, observation protocols (primarily based around 5MBC), data standards, and reporting standards between DOC and regional councils will eventuate should councils adopt the proposed Tier 1 biodiversity indicators.
- There is no agreement that the 5MBC is the preferred protocol for bird monitoring. Different protocols have advantages and disadvantages.
- With the increasing awareness by regional councils on maintaining the provenance and quality of their own data, the data are likely to remain in discrete sources - there is no documented uptake by regional councils of the DOC data management system.
- National-level Tier 1 analyses (and the accuracy of regional-level Tier 1 reports) would be enhanced by having the DOC and regional council bird data used together. A BSS facilitating federation of discrete sources will make it more efficient for researchers to collect up and analyse the data.

Benefits from BSS

- Species-based data-point distribution maps as rough proxies of species distribution[3].
- Species-based data-point distribution maps that show potential gaps in the spatial distribution of the data.
- Discovery system for relevant data from regional councils.
- A data set that can be used to attempt to derive species occupancy, with known limitations of accuracy based on known limitations of the data.
- A data set with known limitations that can be used to test assumptions about birds as trustworthy biodiversity indicators.
- More robust species distribution mapping and modelling, with time series depiction of change, as a consequence of the larger data pool.
- Effortless up-to-date reporting of species occupancy and distribution using data as soon as it is loaded and quality checked.

[2] In the context of the New Zealand Sustainability Dashboard project

[3] While "available today" through GBIF, there is no systematic data capture from all relevant sources, so the picture is not nearly as complete as it could be.

The Domain of Weeds

Case Study

What is the current distribution of 'species x' (of common interest across New Zealand, and to both RCs and DOC)? What is the potential occupancy / scope for further spread of 'species x'?[4]

Current Context

- Regional councils and DOC face the issue of trying to identify and manage pest plants (weeds) strategically – locally eradicating if feasible, and suppressing spread or suppressing negative effects where eradication is not feasible.
- Feasibility of control is supported by an analysis of costs and benefits. There is increasing demand that this analysis is informed by the current national density and distribution of the weed, and on models of the future potential distribution of weeds.
- Knowledge of the dynamically changing national distributions of managed weed species would facilitate national and regional reporting on changes in weed distribution. This information on distribution changes can be used to monitor the success of weed management plans, and to assist in developing models for predicting weed spread.
- The national aggregation of data from nay sources, including regional councils, DOC, collections, and 'citizen science data' is identified as necessary for more effective and meaningful understanding of weed distributions and weed invasions.
- There is no standard system for measuring weed density and distribution. Divergent measurement methods curtail simple "jig sawing" of regional distribution maps. However, there is sufficient consistency to derive national pictures from disjointed datasets, though this requires a high degree of manual intervention. This emphasises the need for national monitoring protocols (which is outside the scope of the BSS project).
- A standards-based federated network capable of providing near real-time national data is preferable to a manual data harvesting, standardisation and aggregation exercise. A manual process is cost effective only when repeated on an infrequent basis.

Benefits from BSS

- Weed data from disparate sources automatically served ready for single-point consumption.
- A data set that can be used to derive species occupancy, with known limitations of accuracy based on known limitations of the data.
- Reduced (costs from) manual transaction of data
- More robust species distribution mapping and modelling, with time series depiction of change, as a consequence of the larger data pool.
- Effortless up-to-date (more timely) reporting of species occupancy and distribution using data as soon as it is loaded and quality checked.
- Fine-scale (sub-regional) tracking of weeds so invasion fronts are better defined, and weed management better targeted (as a consequence of standardising sampling protocols).

[4] In the context of the needs of a proposed National Weeds Distribution Database and the predictive modelling platform under development by AgResearch, 'species x' can be any species regularly monitored as part of a regional council or DOC pest control programme.

The Domain of Fish

Case Study

What is the current distribution pattern of native fish species compared to potential occupancy?

Current Context

- The New Zealand Freshwater Fish Database (NZFFD) hosted by NIWA contains over 30,000 freshwater fish observations. The database is considered to be a “standard”.
- The database has been populated by a number of organisations including NIWA, DOC, Universities, and regional councils.
- Increasing awareness of data collectors to focus on maintaining the provenance and quality of their own data has led councils at least to re-think how they manage their fish data.
- Recent expansion of fish data collection activities may not conform with NZFFD standards which have led to potential contributors holding data separately.
- The quantum non-NZFFD compliant data and the degree of divergence from current guidelines is not known.
- The national aggregation of separate regional council and DOC data, and the NZFFD is identified as necessary for more robust modelling of fish distribution and potential occupancy (as opposed to current models which express “likelihood of capture”).

Benefits from BSS

- Fish data from disparate sources automatically served ready for single-point consumption.
- A data set that can be used to attempt to derive species occupancy, with known limitations of accuracy based on known limitations of the data.
- More robust species distribution mapping and occupancy modelling, with time series depiction of change, as a consequence of the larger data pool.
- Effortless up-to-date reporting of species occupancy and distribution using data as soon as it is loaded and quality checked.

The Domain of Aquatic Macro Invertebrates

Case Study

Demonstrate a way to automatically aggregate regionally disparate macro invertebrate datasets (calculated indices and species). Does the geographic variance in species distribution influence the calculated MCI?

Current Context

- The Macro Invertebrate Community Index (MCI) is a well-known method for establishing a biotic index of stream health and condition and is proposed as one of the next possible indicators for reporting regional council state and trend data at a national level though the Land and Water Aotearoa (LAWA) website.
- However, while it is used by almost all regional councils, the ubiquitous application of the index over the whole of New Zealand (accounting for potential tolerances of species in different bioclimatic zones, landuses, and site-level habitat suitability) has not been thoroughly explored.
- Regional councils generally maintain macro invertebrate records on in-house systems and there is no agreed consistent approach to data management, quality control, and archiving. Some systems may not be machine to machine readable. The extent of consistency is not known.
- Aggregation of invertebrate species records to depict distributions, matched against landuse, habitat, season, and regional differences would facilitate gaining better national context to reporting the MCI index.
- The development of the BSS is congruent with the needs of the LAWA team and collaboration would be beneficial to both parties.

Benefits from BSS

- MCI and raw invertebrate count data from disparate sources automatically served ready for single-point consumption by LAWA and researchers.
- A data set that can be used to attempt to determine species distributions.
- A data set exposing the point-distribution of current and past macro invertebrate sampling regimes that may be used to identify gaps in the national monitoring network.
- Effortless up-to-date reporting of MCI using data as soon as it is loaded and quality checked, with any influence of geographic and protocol variation taken into account.

Interoperable Biodata Infrastructure Strawperson – key features

To facilitate the further discussion on an interoperable biodiata infrastructure and participant readiness, the BSS team identified the following objects to be contained in a BSS code of practice.

Data Management

- Consistent metadata (data-set and methodology)
- Consistent taxa descriptors
- Consistent reference information (geospatial, temporal)
- Consistent attribute descriptors (agreed standard vocabularies)
- Consistent and globally unique dataset identifiers
- Agreed set of quality descriptors (quality codes)

Delivery

- Agreed open standard web services

We drafted as part of a 'BSS strawperson' a number of key minimum fields to be part of a 'BSS code of practice' describing observational records (Appendix 5).

These objects and fields were posed at the workshops and participants were asked to pass comment. The workshop participant make-up was intentionally "biologist" biased (as opposed to "informatics" biased). The result is an agreed minimum data management and exchange protocol that focuses on the "what" (data fields that must be shared and the format of the data). There has been less focus on the "how".

The most important aspects of the agreed 'strawperson' minimum data management and exchange protocol include the following.

- We found general agreement that the proposed fields are based on the base information needs for all use cases (for the biotic domain)[5].
- The proposed fields are almost all Darwin Core / GBIF[6] data compliant or have another pre-existing standard that can be followed.
- The data providers have indicated that the information needed to populate the data elements under consideration for the BSS prototype is available or can be readily sourced.

[5] The base information needs for the habitat and abiotic components of the Case Studies has not been assessed in detail and will require further analysis.

[6] See www.gbif.org

Conclusions and Recommendations

The BSS team believes there is sufficient information from potential data providers to adopt the case studies and to continue progress to BSS Phase Two, namely:

- To develop a draft working 'BSS Code of Practice' (based on our 'strawperson') and guidelines for data publication (by August 2014);
- To work with a few 'data provider agencies' on testing the 'BSS Code of Practise' (by December 2014) for
    - compatibility with the agencies' current data holdings;
    - the agencies' capabilities to implement changes in their internal workflows / data management systems so their data holdings become 'BSS compliant' 'by default'; and
    - the agencies' capabilities to implement data publication mechanisms.
- To work with potential clients / data consumers, specifically the LAWA project and the AgResearch National Weeds project (coordinating with Graeme Bourdot/AgResearch and Richard Bowman/Environment Southland) on testing the 'Code of Practise' to these clients' needs (by March 2015).

All of the key data provider agencies partnering with the BSS project will need to undertake an effective self-review and be prepared for a significant change from current data management practices and workflows to enable data-federation and interoperability. The questionnaires and workshops identified the types of data that project partners have, and their capacity and willingness to change. To help with that, the BSS WIKI (https://teamwork.niwa.co.nz/display/NZBSS) provides some help on how to get 'BSS ready'. In addition, over the rest of Phase One and Phase Two the BSS project team will need to directly help project partners with getting BSS ready including elements like:

- Provide and guide through a standard review tool (the Biodata Guides tool being proposed for this);
- Documentation of barriers that delay the completion of Phase Two and may stall the implementation of Phase Three.

The workshops with regional councils identified an enthusiasm for the project beyond the project partners. There may be willingness for other regions to join the project and contribute to the data provision earlier than Phase Three. The BSS project team should:

- Engage with non-partner councils again and test their willingness to participate; BUT
- Be mindful how much extra work this might put on the BSS team (in terms of the support needed to get BSS ready).

Additional remarks arising from the workshops and discussions which remain rich areas for further thinking include the following.

- The security of sensitive information (particularly endangered species locations) is something the BSS team needs to be cognisant of. It is up to the data provider to ensure data they do not want to be made public stays obscured.
- The ability for the BSS to be able to transmit estimates of data quality is desirable, but it was concluded that quality codes are so dependent on data collection protocol that transmitting fitness for alternate uses is something a BSS is unlikely to be able to serve.

Appendix 1 – Bird Workstream

Proposed Case Studies

The case studies proposed initially were:

- Are common native birds in decline? – assessing if bird such as tui and bellbird are less frequent now than in the recent past;
- Are black petrel populations in decline as a consequence of by-catch? - assessing the by-catch risk in different fisheries;
- Are rooks in decline? - identifying rook source populations and change over time;
- Is there evidence that bird populations are changing in relation to land use intensification? - Examine historic and current distributions and relative abundances of exotic and native bird species relative to landuse change and landuse intensification; and
- Identify candidate species, donor sites, and recipient sites for mainland island translocations.

The most popular choice arising from the questionnaires was the more generic question of whether birds are in decline. Over the course of the workshops, it was established that regional councils were most interested in native bird species distributions and identifying drivers of change. This requirement is congruent with research presently underway to establish birds as an indicator of ecological integrity, as well as work to identify trustworthy biodiversity indicators.

Case Study Question

The BSS project itself is unable to address the question of drivers in changes in bird distributions but consideration of the current and any future bird data to be shared will ensure the BSS can facilitate the answering of such questions. After careful consideration of the data potentially available, and consideration of the information outcome that consumers are seeking, the BSS team has phrased the Bird Workstream research question as;

Can access to dynamic data on bird distributions support the generation of trustworthy indicators for sustainability assessment and reporting?[7]

The relevance of this question as an outcome of BSS (with reference to the information requirements of consumers) sits with the ability to rapidly generate nationally spatially explicit species distribution and occupancy maps running off the back of a large amount of data of known quality.

Existing standards

Respondents and workshop attendees confirmed the relatively common use (or support for) the Winter Garden Bird Survey, 5-minute bird count method, distance methodology, and the OSNZ bird atlas protocol. There are also a few instances among regional councils of specific protocols for specific research questions (e.g. Tui and project Halo in Waikato) or reliance on external parties where the protocol for collection is not known (e.g. use of OSNZ for shorebird surveys in Nelson).

Known issues and shortfalls with existing data collection protocols are well understood, and do not need highlighting for BSS other than to draw attention to the inability of most methods to accurately reflect changes in abundance. Also the most used methods do not measure absence. It will be very important that the BSS is able to carry the data collection protocol along with the data itself. One of the reasons for the current phrasing of the research question, is

[7] In the context of the New Zealand Sustainability Dashboard project

that the protocol used for data collection will influence the fitness of the data. The degree to which data are fit enough for use in national biodiversity indicators reporting needs to be tested.

Some regional councils use existing open data repositories (e.g. e-Bird and NatureWatch) for their data. As potentially standardised data management systems, the common use of these systems could potentially circumvent the need for BSS. However, there is not common agreement among regions to use these systems and even the users of the systems report shortfalls in the way these repositories handle their data.

Availability of data

Relatively few regions report having region-wide or large scale bird monitoring programmes with large amounts of multi-species data to share. Auckland Council's grid-based programme appears to be the largest, followed by Greater Wellington's regional parks and site-led programmes. Other regions report having site-led programmes where some bird data are collected "occasionally" and at least a third of the Councils reported that they do not hold any bird data.

Examples of existing practices for sharing data are limited to those councils such as Greater Wellington that put their data into public repositories, or councils who rely on an external party to collect manage and report the data. By and large it would appear much of the data held by regions is not readily discoverable by a third party.

While there appears to be little data at present, all councils interviewed support the need for a system for data federation to be in place in time for the implementation of the regional council Tier 1 biodiversity indicators programmes. The presumption is that in the near future all regions will be actively collecting (or actively seeking the collection of) bird data as part of the avian indicator for ecological integrity. This is another reason for phrasing the research question as it is. The current availability of data will influence the fitness of species distribution maps – how much needs to be tested. How current bird data gaps can be filled by future monitoring is fundamentally important to regions that may not be able to afford to undertake systematic monitoring of their entire regions, but could afford to fill some critical gaps in data distribution.

BSS Product - Outcomes and Outputs

The following can be anticipated as tangible products that are not available today, but will arise as a consequence of the BSS project:

- Immediate - species-based data-point distribution maps as rough proxies of species distribution[8];
- Immediate - species-based data-point distribution maps that show potential gaps in the spatial distribution of the data;
- Immediate - discovery system for relevant data from regional councils;
- Immediate - a data set that can be used to attempt to derive species occupancy, with known limitations of accuracy based on known limitations of the data;
- Immediate - a data set with known limitations that can be used to test assumptions about birds as trustworthy biodiversity indicators;
- Medium term - more robust species distribution mapping and modelling, with time series depiction of change, as a consequence of the larger data pool;
- Long term - effortless up-to-date reporting of species occupancy and distribution using data as soon as it is loaded and quality checked;

[8] While "available today" through GBIF, there is no systematic data capture from all relevant sources, so the picture is not nearly as complete as it could be.

Appendix 2 – Weeds Workstream

Proposed Case Studies

The case study proposed initially was to:

- Examine if douglas-fir is becoming a nationally significant weed/pest - analysing location data from regional councils and DOC, establish the rate and range of spread.

While questionnaire respondents rated the issue as worth exploring, a number of respondents wanted to see the question widened to include all potential wilding pine species. For those that did not think douglas-fir was a useful line of inquiry, the next most popular species of choice was old man's beard.

Over the course of the workshops it became apparent that few regional councils were presently collecting wilding douglas-fir data and they would struggle to be able to provide data on existing plantations. This effectively renders douglas-fir an obsolete avenue of inquiry to demonstrate the value of a BSS. What also became apparent is that regional councils are highly interested in any species that is out of control in a neighbouring region, but is not widespread in their own. For example with the case of old man's beard, while this species is so widespread in the central South Island that Environment Canterbury manage this species only on a site by site basis, the southern South Island councils have strategies to attempt to have the species under total control (Otago) or to eradicate this pest from their region altogether (Southland).

Case Study Question

After careful consideration of the data potentially available, and consideration of the information outcome that consumers are seeking, the BSS team has phrased the Weeds Workstream research question as;

What is the current distribution of 'species x' (of common interest across New Zealand, and to both RCs and DOC)? What is the potential occupancy / scope for further spread of 'species x'?[9]

The relevance of this question as an outcome of BSS sits with the ability to rapidly amass data from all sources using machine moderated transactions, reducing the transaction time and cost of undertaking this exchange manually.

Existing standards

The regional councils are presently grappling with standards for weed data. Each region has in-house data management practices and individual data needs based around their regional pest management plans. Plans are highly region-centric and as a result there is regional divergence and diversity in how species abundance is measured and classified. Until very recently, commonality and standardisation of data practice among the regional councils has been exceptional rather than normal. However, as a collective of pest management agencies, regional councils have become aware of the need for common practices and the value of being able to share data about pest species distributions between themselves and with DOC. The most tangible evidence of the attempt to find commonality is the evolution of the proposed National Weeds Distribution Database[10].

However, while standards for weeds data are emerging, sharing of data remains a highly manual process. One of the reasons for phrasing the research question as it is, is that the BSS project seeks to use existing standards rather than run roughshod over emerging good practice. The question posed aligns the BSS project with the existing manual

[9] In the context of the needs of a proposed National Weeds Distribution Database and the predictive modelling platform under development by AgResearch (Graeme Bourdot, AgResearch, pers. comm.), 'species x' can be any species regularly monitored as part of a regional council or DOC pest control programme.

[10] Graeme Bourdot, AgResearch, pers. comm.

transaction occurring, with the view of enhancing that with additional standards and practices that will facilitate machine to machine transaction, reducing the transaction cost.

Availability of data

Most regional councils have some form of spatially explicit data for the pests they manage under their respective regional pest management plans. However, the distribution, accuracy, and frequency of repeat observations vary dependant on how the pest is being managed. For instance, Horizons RC has both point and polygon data for old man's beard observations, repeated annually in areas where the pest is actively managed. However, there is a zone where active management is not undertaken because the pest is deemed to be too widespread. The density and distribution of individual infestations within that zone are not regularly assessed. A regional council may choose not to have a pest plan for weeds that are widespread or species that are so sparse that they are believed not to be a threat. In either case, the collection of data for such species will be negligible, creating a potentially false sense of absence. For species that are actively being managed, most regions measure some form of direct or coded abundance from which a sense of presence can be obtained.

Concurrent to the BSS project is work by AgResearch to manually collate presence data from DOC and the regional councils and use this to derive national weed distribution maps and use the data for predictive modelling[11]. The need for BSS lies with an enhanced system for data federation that reduces the level of manual transaction and increases the speed that species occupancy maps can be refreshed.

As more effort goes toward defining common data management practices, an actively evolving BSS may encourage practitioners to standardise the way weed abundance is measured and reported at site-scale. This could enhance the accuracy of weed predictive mapping as it eliminates the data collection protocol as a source of variation.

BSS Product - Outcomes and Outputs

The following can be anticipated as tangible products that are not available today, but will arise as a consequence of the BSS project:

- Immediate - weed data from disparate sources automatically served ready for single-point consumption;
- Immediate - a data set that can be used to attempt to derive species occupancy, with known limitations of accuracy based on known limitations of the data;
- Immediate – reduced effort and manual transaction of data;
- Medium term - more robust species distribution mapping and occupancy modelling, with time series depiction of change, as a consequence of the larger data pool;
- Long term - effortless up-to-date reporting of species occupancy and distribution using data as soon as it is loaded and quality checked;
- Long term – fine-scale (sub-regional) tracking of weeds so invasion fronts are better defined, and weed management better targeted (as a consequence of standardising sampling protocols).

[11] Graeme Bourdot, AgResearch, pers. comm.

Appendix 3 – Fish Workstream

Proposed Case Studies

The case studies proposed with the questionnaire were (paraphrased) to:

- Assess if land use change is the driver for reduction in native fish populations and diversity - use native fish presence/absence data and land use data, and test correlations;
- Assess habitat for trout including food supply, access, spawning habitat etc. - use presence/absence data to assess habitat and food supply suitability;
- Assess the pest potential of a (to be named) exotic fish – identify effects and spread
- Assess the potential of a (to be named) fish pathogen – identify effects and spread

The most popular choice from questionnaire respondents was the issue o identifying if landuse change is the driver for reduction in native fish population and diversity. A number of respondents however commented that any effort to better depict native species distributions was first and foremost. During the workshops it became even more apparent that most users of fish data are interested in questions of “where is it?” and “where else does it occur?”.

Case Study Question

After careful consideration of the data available, and the more general nature of the query that participants expressed, the BSS team has phrased the Fish Workstream research question as;

What is the current distribution pattern of native fish species compared to potential occupancy?

The relevance of this question as an outcome of BSS sits with the ability to amass fish presence, abundance, and habitat data from all sources.

Existing standards

Workshop participants who reported having collected fish data stated either that their collections were sporadic based on resource consent requirements, or that they followed “more or less” the “Standardised Fish Monitoring for Wadeable Streams” (David and Hamer) protocol. By “more or less” some respondents stated that they used the protocol but had made some slight modifications (extension of time or reach length fished) in order to extend effort. To this extent, while practices are not fully “standardised”, data collectors are following knowable and repeatable processes. It will be important that search effort is carried with the observation.

Councils that collect fish data use the NZ Freshwater Fish Database (NZFFD) to the extent that they can lodge the observation following the standards set for that data repository. Councils that use the David and Hamer protocol now tend to manage the data in-house using excel spread sheets specifically designed to hold all of the data collected using the protocol.

One of the reasons for phrasing the research question as simply as it is, is that the main issue the BSS project needs to solve in the first instance is how to amass existing data. Occupancy is a secondary research question that can only be answered once the data are federated.

Availability of data

The regional councils following the David and Hamer protocol report that manual transaction of the data will be relatively simple. Data that have been collected as part of a resource consenting process will be more difficult to source as these tend to be filed by consent-relevant details and not by biotic-relevant details. Resource consent data may have been exported to the NZFFD as a consequence of the first-party contractors (consultants for the consent holder) but the degree to which this is common practice across New Zealand is not known.

When asked if the practice of extending search effort was akin to identifying absence, respondents were of mixed views. Some felt that because a particular fish species was not always targeted for survey, then true absence can not be verified. However, where intensive search effort has been put in to find a particular species to no avail, absence can be presumed. Across all of the respondents who are collecting fish data, all agreed it is possible to derive presence from their data because they will either have direct or coded abundance as part of the record.

The need for BSS lies with an enhanced system for data federation that exposes the data collected by regional councils.   As more effort goes toward defining common data management practices, an actively evolving BSS may encourage more practitioners to adopt current guidelines as standardised sampling protocols.

BSS Product - Outcomes and Outputs

The following can be anticipated as tangible products that are not available today, but will arise as a consequence of the BSS project:

- Immediate - fish data from disparate sources automatically served ready for single-point consumption;
- Immediate - a data set that can be used to attempt to derive species occupancy, with known limitations of accuracy based on known limitations of the data;
- Medium term - more robust species distribution mapping and occupancy modelling, with time series depiction of change, as a consequence of the larger data pool;
- Long term - effortless up-to-date reporting of species occupancy and distribution using data as soon as it is loaded and quality checked;

Appendix 4 – Aquatic Macro Invertebrate Workstream

Proposed Case Studies

The case studies proposed with the questionnaire were (paraphrased) to:

- Aggregate macro invertebrate community Index (MCI) scores - collate the MCI scores for permanent monitoring sites from regional and NIWA networks to produce site-level MCI trend statistics to be reported via the Land Air and Water Aotearoa website (LAWA).
- Examine invertebrate presence data to detect if there are regional variations in the MCI that is derived as a consequence of season, co-dependant on water quality, landuse, and disturbance regimes.
- Identifying gaps in monitoring network - use national and local invertebrate data to identify gaps (the type of gap and the cause thereof) in the monitoring network in an area of interest.
- Examine the suitability of restored terrestrial (riparian) habitats – are restored habitats suitable to the needs of aquatic macro invertebrate adult life stages?

The most popular choice from questionnaire respondents was the ability to automatically deliver MCI scores to LAWA. The result was not too surprising as all councils have recently been instructed by LAWA to prepare their MCI datasets ready for consumption by this national environmental information reporting portal. At the workshop, the demand was less clear cut with participants also expressing strong interest in the ability to re-use the raw macro invertebrate (species) data to explore other questions such as regional distributions of species and the effect of biogeographical variation on the derived MCI score.

Case Study Question

After consideration of the influence biogeography may have on the MCI score, the BSS team has phrased the Macro Invertebrate Worksktream research question as;

Demonstrate a way to automatically aggregate regionally disparate macro invertebrate datasets (calculated indices and species). Does the geographic variance in species distribution influence the calculated MCI?

The relevance of this question as an outcome of BSS sits with the ability to amass both the calculated MCI scores (i.e. the MCI:value is exposed the same as species:count), and the raw species data.

Existing standards

Almost all of the regions interviewed collect aquatic macro invertebrate data as part of their normal State of the Environment Monitoring programmes. These samples tend to be collected, sorted, and analysed (by the calculation of the Macro Invertebrate Community Index (MCI)), following published protocols. There is a choice of protocol for sample collection, sorting, and calculation as a consequence of differences of intended purpose for the data (whether primarily qualitative State of the Environment reporting, or primarily quantitative effects-based reporting for (say) resource consents). Horizons RC experience is that choice of sampling protocol has some bearing on the reported MCI score and can alter the conclusion regarding the state of the water or habitat quality.

One of the reasons for phrasing the research question to design a product that is able to expose both the calculated MCI and the underlying raw invertebrate data is because protocol variance and regional biogeographical variance may be strong influencers of the reported MCI. Councils want to understand the effects of these variances better to better identify if regional comparisons of MCI via LAWA are a fair reflection of differences in water quality, and not an artifice of sampling or sorting method, or biogeographic differences.

Availability of data

Participants reported that they hold MCI and raw invertebrate values either in excel spread sheets or on custom (in-house) databases. At least five of the councils (Environment Southland, Nelson, Greater Wellington, Horizons, and

Hawkes Bay) have the cut down version of the Cawthron Archival Data and Delivery Information System (CADDIS – fly), though this is not being used extensively or consistently for archiving aquatic macro invertebrate data. This database is able to store the raw invertebrate data and the MCI as an analyte in an exposable Access or SQL database. The five councils mentioned have begun work together to be more consistent with their use of the database. Most other councils report they have legacy systems that they believe will be able to be tapped to readily expose the macro invertebrate data. The councils are preparing to serve other water quality information for automated ingestion into LAWA. Confidence is high that the same can be done for aquatic macro invertebrate data.

Participants raised concern that variation in geography and choice of MCI calculation method has some influence on the reported MCI, and therefore the MCIs reported in LAWA may misconstrue the state or the water quality or habitat. Councils expressed strong desire for rapid federation of the raw invertebrate values so that questions about species distribution and biogeography might be readily answered.

BSS Product - Outcomes and Outputs

The following can be anticipated as tangible products that are not available today, but will arise as a consequence of the BSS project:

- Immediate - MCI and raw invertebrate count data from disparate sources automatically served ready for single-point consumption by LAWA and researchers;
- Immediate - a data set that can be used to attempt to determine species distributions;
- Immediate - a data set exposing the point-distribution of current and past macro invertebrate sampling regimes that may be used to identify gaps in the national monitoring network;
- Long term - effortless up-to-date reporting of MCI using data as soon as it is loaded and quality checked, with any influence of geographic and protocol variation taken into account;

Appendix 5 - Strawman BSS: minimum data management and exchange protocol

The following table lists data elements currently under consideration for the BSS prototype (see also BSS WIKI).

Strawman BSS: data format and field name protocol						
category	**data field**	**DWC (non-DWC)**	**notes**	**format**	**development requirements**	**reference**
spatial	point	decimalLongitude, decimalLatitude, verbat	polygon if recorded	OGC simple features (two values+ coordinate syst	specification	OGC
	polygon	footprintWKT, footprintSRS	central point of observations if recorded		specification	OGC
	line	???	observations along a path (e.g. for birds and fish)		specification	OGC
	observed_area	observedArea	estimated observed area if recorded; note: qualit	integer		
	site identifier	locationID	simple text dealt by institutionally and guided by	text string, guided by best practice	best practice guideline needs developme	
	spatial accuracy qualifier	???	text according to vocab	vocab: gps, manual estimate from map, just rough	vocab/convention	
	NZ region	stateProvince		text string according to vocab	convention	LINZ Gazet
	locality	verbatimLocality		text string according to vocab	convention	LINZ Gazet
	ref for location definition	*locationAccordingTo*	*fixed according to convention*	*http://www.linz.govt.nz/placenames/find-names/*	*convention*	
temporal	date/time (interval)	eventDate	incl. duration /time interval options!	ISO8602	specification	ISO8602
species	observed_species	taxonID	national vocabulary	vocab service	vocab service	NZOR
	species name as textfield	scientificName		text string		
	uncertainty in species identification	identificationVerificationStatus	vocabulary or convention: verified (standard), un	text string according to vocab	vocab/convention	
observation	abundance_category	occurrenceStatus	national vocabulary or convention	vocab: rare, common, present, observed absent	vocab/convention	
	individual_count, mincount,maxcou	individualCount, minCount, maxCount	count of observed individuals	integer		
	% coverage, min%,max%	percentCoverage,minCoverage,maxCo	% coverage over observed area	integer		
	covered_area, min, max	areaCovarge,minArea,maxArea	total covered area	integer		
	uncertainty of abundance	occurrenceUncertainty		vocab: accurate, estimate, guess	vocab/convention	
other observatior	lifestage	lifeStage	lifestage of observed organisms	vocab: baby, child, adult	vocab/convention	
	reproductive status	reproductiveCondition	reproductive status of observed organisms	binary	vocab/convention	
	sex	sex	sex of observed organisms	male/ female/ other	vocab/convention	
	length, min, max	???	length (range) of observed organisms	integer	???	
	wild/not wild	establishmentMeans	not sure if useful, maybe too confusing	vocab	vocab/convention	
	related biotic factor (?)	???	eg eel cyst, fish parasite, kauri dieback, relates to	text string	???	
metadata	observer	identifiedBy	one or many observer(s)	text		
	owner/author?	dcterms:rightsHolder??	dataset owner	text		
	institution	institutionCode		text	best practice according to some business	
	title of dataset	datasetName		text		
	description of dataset	[in service metadata??]		text		
	method	samplingProtocol		vocab/text	how to do vocab; at all possible??	
	methodreference	???	citation	text string	best practice??	
	eventremarks	eventRemarks	any comments with regards to the observations	text		
	uid for each observation ??	occurrenceID	need to define a standard here	some best practise with organisation code or regis	???	

Appendix 6 - Getting BSS ready

The following table lists some fundamental BSS requirements, and ways data providers can prepare for sharing of bio data through BSS (see also BSS WIKI).

BSS Requirements	Actions / Questions for data providers
Species abundance descriptors need to be standardized nationally through agreed vocabularies	Check internal usage of abundance descriptors and categories with a view to have a institutional standard
Taxa names need to be standardized nationally through a national register	Develop an in-house standard taxa list everybody needs to use and is in all workflows Get ready to link to New Zealand Organisms Register: add nzor_id in your in-house taxa list. Have a look at nzor.org.nz, demo.nzor.org.nz Goto http://demo.nzor.org.nz/matches, there you can match your taxa list against NZOR!
All spatial coordinates for a occurrence observation need to be described consistently and unambiguously	Are you sure you know your coordinate system used for all observations?
All dates / times for an occurrence observation need to be described consistently and unambiguously	Are you sure all your date / time recorded are unambiguous (time zones, daylight saving)?
Some records require site identifiers	How do you record site identifiers for occurrence observation at the moment? Do you have a well maintained site register?
General data management	Do you know where all your bio observation data is? Consider establishing a data catalogue Consider establishing an institutional bio-data archive Consider doing a maturity assessment for your data management system, using the Biodata Management Guide developed through Dataversity: http://dataversity.org.nz/guide/