Manaaki Whenua
Landcare Research

A Review Of Major New Zealand Biodatabases' Readiness To Integrate With GBIF

Prepared for the Terrestrial and Freshwater Biodiversity Information System Fund by Jerry Cooper and Julian Carver

3 June 2005

Table of Contents

1 Executive Summary

New Zealand has significant amounts of biodata[1] stored in databases held by research institutes, universities and museums. This sort of biodata is becoming increasingly valuable in understanding changes in biodiversity at local, national, and global levels. In February 2001 New Zealand signed a memorandum of understanding to join the Global Biodiversity Information Facility (GBIF). This facility provides a way of connecting up biodiversity data internationally. To date only a small percentage of New Zealand's biodiversity data has been connected up to GBIF, and this was done only as a once off 'exemplar' project.

This report documents a survey conducted on behalf of the Terrestrial and Freshwater Biodiversity Information System (TFBIS) fund to determine how ready major New Zealand bio databases are to integrate with GBIF (the Global Biodiversity Information Facility).

Major biodatabases surveyed included those held by The Department of Conservation, Te Papa, Canterbury Museum, Otago Museum, Lincoln University, Massey University, Forest Research, NIWA, Landcare Research, and Cawthron Institute were surveyed. Other universities and major museums were contacted but no responses were received.

33 datasets in total were surveyed, half of which were collections and half observations. There were around 800,000 collection items and 2.8 million observation records. There were a wide range of organism types represented including plants, vertebrates (mammals, birds, fish, reptiles), invertebrates (spiders and insects), and freshwater macroinvertebrates.

Respondents were asked a number of questions about ability to export data to GBIF, data accuracy/ quality, network and data connectivity, legal and intellectual property issues, and organisational willingness. The large major of datasets were able to easily export the basic minimum fields required by GBIF. Many were also able to relatively easily export quite a number of additional fields. Museums were able to export many more fields than the other organisations surveyed.

For the great majority of datasets surveyed there were enough data quality measures in place for organisations to be reasonably confident that their data was of high enough quality to integrate with GBIF. This was however quite a subjective issue, and in some cases could still pose barriers to GBIF integration.

Respondents were asked to rate their ability to connect their data to GBIF in terms of bandwidth and network connectivity. While universities tended to rank themselves low on this scale, the very large majority of collections and observations are held in organisations that could provide more than adequate network connectivity for a GBIF node. Almost all the data could be updated to GBIF at least a few times a year. Around half the data could be updated on at least a monthly basis. Museums were more likely than others to be able to update their data on a daily basis as they have single large collection management systems which they intend to make accessible over the web to the general public within the next year.

For organisations surveyed it appeared that a reasonable proportion of the data they could put into GBIF was not encumbered by any legal or intellectual property constraints. Universities were least likely to have these sorts of constraints, followed by museums, and then research institutes. The lower ranking for research institutes was due primarily to the fact they were holding data on other people's behalf (especially true for observations), or due to safety issues (biosecurity, rare or threatened plants). The issue of scientist's publishing rights and intellectual property while still important appeared secondary to these issues.

The large majority of the biodata was in organisations that said they would like to contribute data to the GBIF network, and would have executive support for doing so. Funding was the primary issue

[1] See Glossary in section 5.2, page 32 for an explanation of this and other scientific/technical terms used in the document

here. Almost all the data is held by organisations who said they may be able to fund connecting to GBIF to a limited level themselves, but definitely would be willing to do more if there was external funding available. It appears essentially that much of New Zealand's biodata is in a state where it could be connected to GBIF, but unsurprisingly perhaps this seems unlikely to happen without specific funding to do so.

Based on the analysis and conclusions a number of recommendations are made for consideration by the TFBIS committee. These include focusing on the 'low hanging fruit' collections data at the major museums, finding ways to encourage universities to improve their biodata management, balancing connection to GBIF with ongoing digitisation efforts, and considering the additional benefits in terms of awareness of New Zealand biodata that may be created by implementing a national GBIF portal.

2 Introduction

2.1 Purpose of the Report

This document reports on the results of a survey conducted on the readiness of major New Zealand bio databases to integrate with GBIF (the Global Biodiversity Information Facility).

The management committee for the Terrestrial and Freshwater Biodiversity Information System (TFBIS) fund wished to investigate some aspects of the status of bio databases in New Zealand. The purpose of the investigation was to provide the TFBIS committee with information to support its strategic planning and future funding decisions.

A similar investigation was run at the same time into the status of regional council bio databases, and into issues associated with taxonomic names and associated databases. These are documented in separate reports.

2.2 Context

2.2.1 What TFBIS is

The Terrestrial and Freshwater Biodiversity Information System (TFBIS) Programme supports the conservation of New Zealand's indigenous biodiversity, by increasing awareness of and access to fundamental data and information about terrestrial and freshwater biota and biodiversity. The Programme is one of a suite of new programmes that reflects Government's commitment to achieving the goals of the New Zealand Biodiversity Strategy (NZBS). The background to TFBIS is as follows.

In February 2000 the Government adopted the New Zealand Biodiversity Strategy (NZBS) to halt the decline in the variety of naturally occurring plants, animals and ecosystems in New Zealand.

In June 2000, the Government announced a Funding Package of $187 million over five years to achieve the goals of the NZBS. This funding has enabled biodiversity management agencies to increase their 'hands on' work programmes, e.g. to manage more threatened species and a wider range of ecosystems, and to initiate other new work. The Terrestrial and Freshwater Biodiversity Information System (TFBIS) Programme has been allocated $9.6 million (GST inclusive) over five years and $2.714 million annually thereafter.

The Department of Conservation (DOC) administrates the TFBIS Programme, but it is the Department's view that the Programme is for the benefit of all agencies and organisations that contribute to the management of New Zealand's indigenous biodiversity. More information on TFBIS can be found at http://www.biodiversity.govt.nz/land/nzbs/tfbis/tfbis/index.html.

2.2.2 What GBIF is

Biodiversity informatics is a branch of computer science dealing with information about living organisms. The Global Biodiversity Information Facility (GBIF) exists to make the world's biodiversity data freely and universally available by developing biodiversity informatics tools to provide web access to primary information on the world's organisms.

GBIF's particular focus is on data about species and about individual specimens representing those species, although appropriate links will also be made to relevant eco-logical and genetic data. Even within the more restricted domain of species and specimen data, the range of information is enormous and the data are currently held in hundreds of differing formats.

The following is a list of the biodiversity data that GBIF wishes to make available:

1. Taxon names
2. Taxon occurrence information (primarily species-level, but including data for taxa at different ranks where appropriate) including specimen records (from natural history collections) and observation records
3. Links to other information, including: taxon descriptions, information on taxon biology and life history, ecological interactions, genetic data, sound and image resources

GBIF is a distributed query network comprising data provider nodes all around the world. More information on GBIF can be found at http://www.gbif.org.

New Zealand signed the GBIF Memorandum Of Understanding in February 2001 and became a GBIF participant. The New Zealand Ministry of Research Science & Technology (MRST) provide annual funding to support a GBIF participant Node Manager (currently about 0.1 FTE). This responsibility is currently undertaken by Jerry Cooper of Landcare Research. In addition MoRST and TFBIS paid for a one-off example GBIF data load of 1.4 million specimen/observation records from the Nationally Significant Databases and Collections at Landcare Research. This is currently the only known GBIF connected data in New Zealand.

More information on the New Zealand GBIF node can be found at http://www.gbif.org.nz.

2.3 Survey Scope

For the purposes of this survey 'major biodatabases' were defined as those involving terrestrial and/ or freshwater biodata as held by The Department of Conservation, Te Papa, other major museums, Universities, NIWA, Landcare Research, Forest Research, and Cawthron Institute. An emphasis was placed on databases containing primary species specimen or observation records. The term 'major' was left deliberately open so as to a) attempt to cover the largest and most important datasets and b) to not exclude new or smaller datasets that could be of interest to GBIF and would give the TFBIS committee a view on the 'state of the play' across all biodatabases.

Specifically out of scope for the report were bio databases held by other private research institutions, NGOs (e.g. QEII Trust, NZERN), and secondary bio data records held by other government departments (such as MAF, MfE, MFish).

2.4 Survey Process

The survey involved:

- A definition of what 'readiness to integrate with GBIF' means, including a set of criteria for assessing readiness
- An assessment of the reported status of the major bio databases against these 'readiness' criteria

The assessments were done using a combination of written responses, phone interviews, and face-to-face interviews with one or more representatives from each of the organisations responsible for the databases.

The survey did not involve a formal 'audit' of the databases involving physically checking the databases for achievement against assessment criteria. The assessments were based purely on written or verbal responses from participants.

2.5 Responses

Responses were received from:

- Cawthron Institute – Karen Shearer, Biologist; Paul Barter
- Forest Research – Chris Ecroyd, Herbarium Curator
- Landcare Research – Sue Sheele, Susan Wiser, Eric Spurr, Aaron Wilton, Wayne Fraser, Peter Johnston, Shaun Pennycook, Trevor Crosby, Jerry Cooper, Nick Spencer
- NIWA – Don Robertson, General Manager, Biodiversity, Biosecurity & Information Services; Steve Massey, Data Management Architect
- Department of Conservation – Tony Charles, Applications Development Manager; Malcolm Harrison, Senior Systems Analyst; Jim Lynch, NHMS Programme Manager
- Canterbury Museum – Paul Scofield, Curator of Vertebrate Zoology
- Otago Museum – Brian Patrick
- Te Papa – Patrick Brownsy, Senior Curator Natural Environment; Philip Edgar, Collections Information Manager
- Lincoln University – Karen Armstrong, Project Leader Molecular Diagnostics
- Massey University – Jill Rapson, Plant Ecologist

Auckland Museum, University of Waikato, and Otago University were approached about the survey but either contact could not be made within the timeframes required, or the survey was sent and a response was not received.

3 Conclusions

From analysis of survey results a number of conclusions can be drawn. These are followed by some recommendations for consideration by the TFBIS committee. These conclusions depend on the analyses in the summary of findings in section 5.4 but are included at the beginning of the document for those that may not wish to review the findings in detail. Conclusions are:

- Collections data is generally higher quality than observations data and would be easier to connect to GBIF for this and other reasons including data standards adherence, fewer legal/IP issues and in most cases better data connectivity.
- The larger museums are the 'readiest' to connect their data to GBIF, followed by DOC, then research institutes, with universities trailing further behind. It should be noted that DOC's readiness is dependent on a fairly large and complex system currently under development.
- Universities seem to have a much lower level of biodiversity informatics infrastructure than CRIs or Museums. CRIs have more complicated and sophisticated informatics infrastructure as they deal with many different kinds of datasets, including observations as well as collections. Museums tend only to deal with collections, and have much less informatics infrastructural complexity as they generally have all their data in one large collections management system.
- The majority of datasets surveyed have defined certainty statements of some kind, which means data quality may not be a significant barrier to integration of New Zealand's major biodatabases with GBIF. Given the subjectivity of responses to this type of question however there may be some cases where it is a barrier. It seems likely that even given stated quality levels additional work would be required in many cases to get data ready to connect to GBIF. More in-depth investigation would be required to confirm this, however this is perhaps better identified by funding bid processes for organisations wishing to connect their data to GBIF.
- Connectivity constraints are far more about data connectivity than network connectivity. More than 99% of collections/observations are in systems held by organisations that could host their own GBIF connected node. Organisations holding the majority of biodata have the bandwidth and network capacity, but are less sure about their ability to keep the GBIF connected data up to date from primary sources.
- Level of willingness and capacity to connect data to GBIF is almost exclusively a funding issue. Managers and scientists would be willing in principal but most organisations are unlikely to fund this themselves. This is especially true of universities and research institutes, and less true of museums surveyed, although for museums funding is likely to make it happen faster.

By way of summary connecting New Zealand's biodata to GBIF is something that by and large we are ready to do, want to do, have agreed to do as part of an international agreement, but just have not yet done in a systematic or committed way. The direct benefits of integrating data with GBIF are difficult to predict. When benefits do come they may well be diffuse, distributed, and perhaps even hard to trace back to the action of connecting to GBIF.

GBIF however is a potential 'attractor' for metadata, i.e. something that will collect metadata about biodatabases, not for its own sake, but as a side effect of achieving another end (in this case the ability to combine and interpret data globally). This 'side effect' may have significant other benefits for New Zealand in terms of increasing awareness nationally about the research data we have. This increased awareness and accessibility could lead to better uptake of research by

biodiversity and resource managers. It may also lead to better research being done, more reuse of existing data, and an increased ability to attract international researchers and research funds.

3.1 Recommendations

Based on these conclusions, and other findings from the analysis a number of recommendations are made. These are simply ideas for discussion and consideration by the TFBIS committee rather than fully formed proposals.

1) Museums, in particular Te Papa and Otago seem the most 'ready' to connect to GBIF. Collections are the 'underpinning' or foundations of biodiversity research so this may be a good place to start. Consider funding efforts to connect these museums once their collection management system and web access projects are completed.
2) Find ways to encourage universities to improve their basic biodiversity information management infrastructure so that some time in the future their data too can be connected globally
3) Think carefully about the balance between digitising more of existing records and connecting what we already have digitised globally. One should not happen at the expense of the other.
4) Ensure that any efforts to connect data to GBIF take into account the fact that separating data from its surrounding information, and the knowledge in the heads of the people that collected it can be dangerous. When taken out of context misinterpretation becomes far more likely. Find ways to retain the transfer of knowledge and insight as we move into an age where data is increasingly freed up. At the same time use initiatives like GBIF to raise the bar in terms of data quality standards in New Zealand.
5) Consider a local GBIF 'harvester' or portal. This could index and describe just New Zealand's biodata. This could have significant benefits in terms of increasing awareness about our biodatabases and the research that has been done around them, and may lead to greater uptake of research findings. It may encourage greater inputting of data by resource/biodiversity management agencies, and could lead to less reinvention of the wheel nationally, in terms of system development, data gathering, and research done.
6) Consider the commonalities and differences between a New Zealand GBIF node/portal, and NIWA's Freshwater Biodata Information System (FBIS) model. There may be synergies in terms of data gathering and other aspects. What role might GBIF play in moving New Zealand closer to having a TBIS (where T is 'terrestrial'), even if it is a 'meta-system' involving many individual distributed systems?

4 Summary of Findings

The survey asked questions about:

- Metadata – database title, custodian, abstract, spatial and temporal extent, and numbers of specimens/observations and species
- Data export – the ease to which data could be exported using GBIF exchange schemas
- Accuracy/Quality – the quality of the data, measures used to document/ensure quality, and whether external or internal data dictionaries were used as quality standards
- Connectivity – the physical network connectivity the organisation could provide between its data and GBIF, and how often the data could be updated to GBIF from primary repositories
- Legal/IP – what proportion of the data might be not available due to legal, intellectual property, or safety reasons
- Willingness – the organisation's willingness and capability to connect data to GBIF

Findings are summarised in the remainder of this section, and detailed response records are provided in the Appendix in section 5.2.

4.1 Datasets Surveyed

33 datasets in total were surveyed. These were:

Canterbury Museum
- Invertebrate Zoology Collection
- Vertebrate Zoology Collection

Cawthron
- Cawthron Macroinvertebrate Data

Department of Conservation
- Bioweb Bird banding
- Bioweb Herpetofauna
- Bioweb Threatened plants
- Bioweb Weeds

Forest Research
- Forest Research Herbarium

Landcare Research
- 5 Minute Bird Counts Database
- Allen Herbarium Specimen Database
- International Collection of Micro-organisms from Plants and Associated Databases (ICMP)
- Mammal Distribution Database
- National Vegetation Survey Databank (NVS)
- New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC)
- New Zealand Fungal Herbarium and Associated Database
- Plant Names Database

Lincoln University
- Centre of Research Excellence Database

Massey University
- Dame Ella Campbell Herbarium (MPN) Database

NIWA
- FBIS Algae Data
- FBIS Aquatic Weed Data
- FBIS Benthic 2000 Data
- FBIS Lake Data Macrophytes
- FBIS Lake Water Quality Data (Algae)
- FBIS New Zealand Freshwater Fish Data
- FBIS Stream Invertebrates Data

Otago Museum
- Natural Environment Collection

Te Papa
- Arthropod Collection
- Birds Collection
- Fishes Collection
- Land Mammals Collection
- Molluscs Collection
- Natural Environment Collection
- Plant Collection
- Reptiles & Amphibians Collection

A number of other datasets were noted during the survey process but were not analysed in any depth because they were too small to be considered 'major' biodatabases. This list is included in the appendix (section 5.2).

4.1.1 Making Sense of the Datasets

Of the datasets surveyed 16 were collections, 16 were observation records, and 1 was a list of taxonomic names.

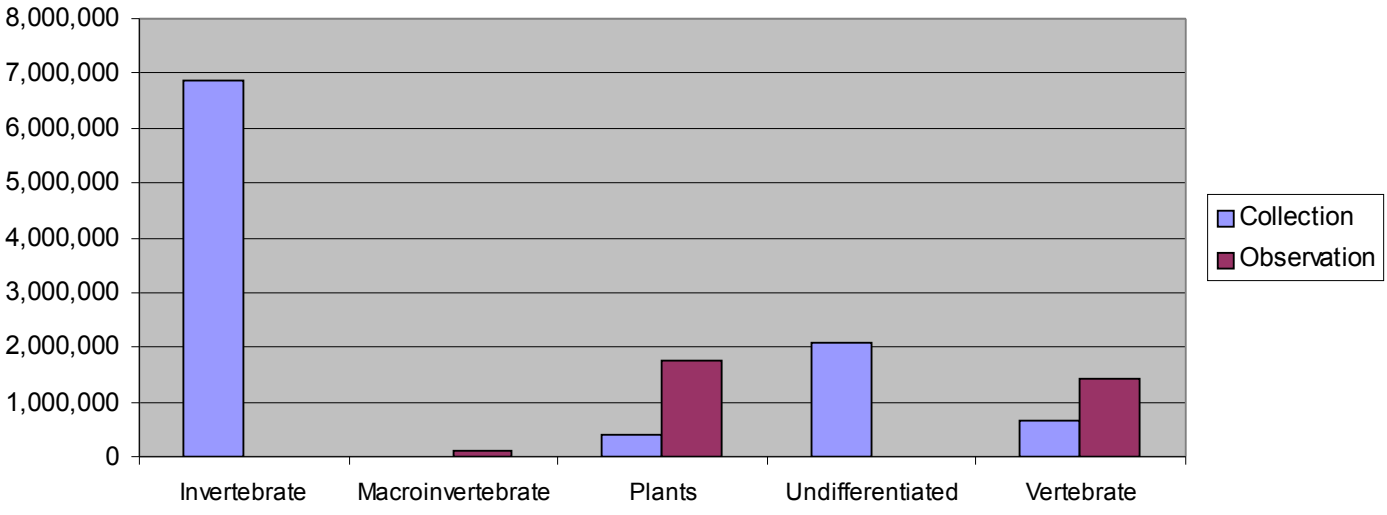
Respondents were asked for the total numbers of collections/observations, the total they had in the database, and the number of taxa represented in each collection or observation dataset. Useful figures were available for the majority of datasets. Complications only arose with TePapa's molluscs data which was only numbered in lots (110,000), with each lot containing from a few to many collections, and with FBIS periphyton data which was measured in number of sites (970). For the purposes of statistical analysis and comparison the periphyton data was ignored, and the undigitised mollusc data left out.

In total there were 10 million collection items in collections surveyed. There were 3.3 million observation records. When broken down by organism type, numbers were as follows. Those 'undifferentiated' were from the Otago Museum Natural History Collection, for which a breakdown by type of organism was not available.

	Plants	Vertebrate	Macroinvert.	Invertebrate	undifferentiated	Total
Collections	387,000	670,500		6,850,000	2,100,000	10,007,500
Observations	1,766,040	1,430,266	125,100	3,000		3,324,406
Total	**2,153,040**	**2,100,766**	**125,100**	**6,853,000**	**2,100,000**	**13,331,906**

For the purposes of this analysis a distinction was made between collections/observations that were described as containing invertebrates and those containing macroinvertebrates. Where macroinvertebrates were referred to they were all in the context of freshwater datasets. Invertebrates were all in relation to terrestrial datasets. The only exception was the Lincoln University Centre of Research Excellence Database which contained both terrestrial and acquatic insects. The numbers in this database were so low in proportion to others that they were unlikely to skew results significantly.

The total number of collection and observation items can be seen as follows:

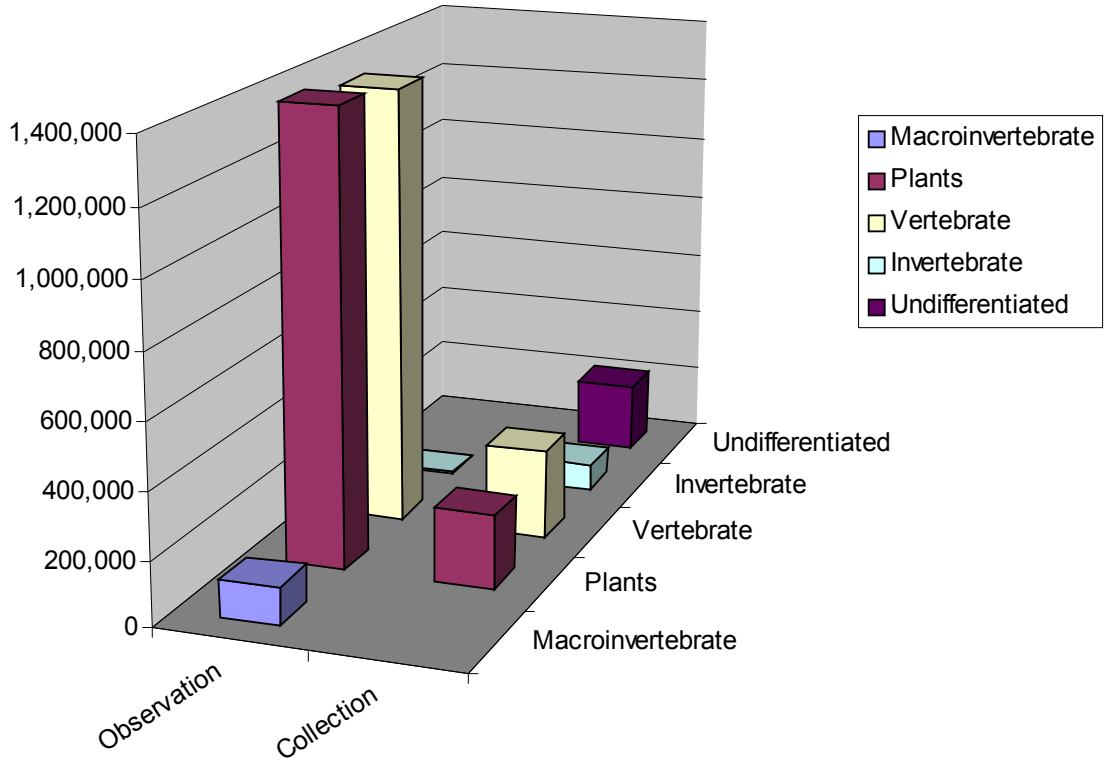


Only a proportion of collections and observation records are currently digitised. In terms of total numbers these were:

	Plants	Vertebrate	Macroinvert.	Invertebrate	undifferentiated	Total
Collections	225,000	273,500		84,000	210,000	792,500
Observations	1,388,040	1,352,366	115,100	3,000		2,858,506
Total	**1,613,040**	**1,625,866**	**115,100**	**87,000**	**210,000**	**3,651,006**

This is represented visually on the following page.

A breakdown of percentages digitised is also included in the subsection below. For the purposes of analysis of survey responses only the figures for the proportion digitised (i.e. recorded in the



databases) were used.

Number of Collections/Observations by Organism Type

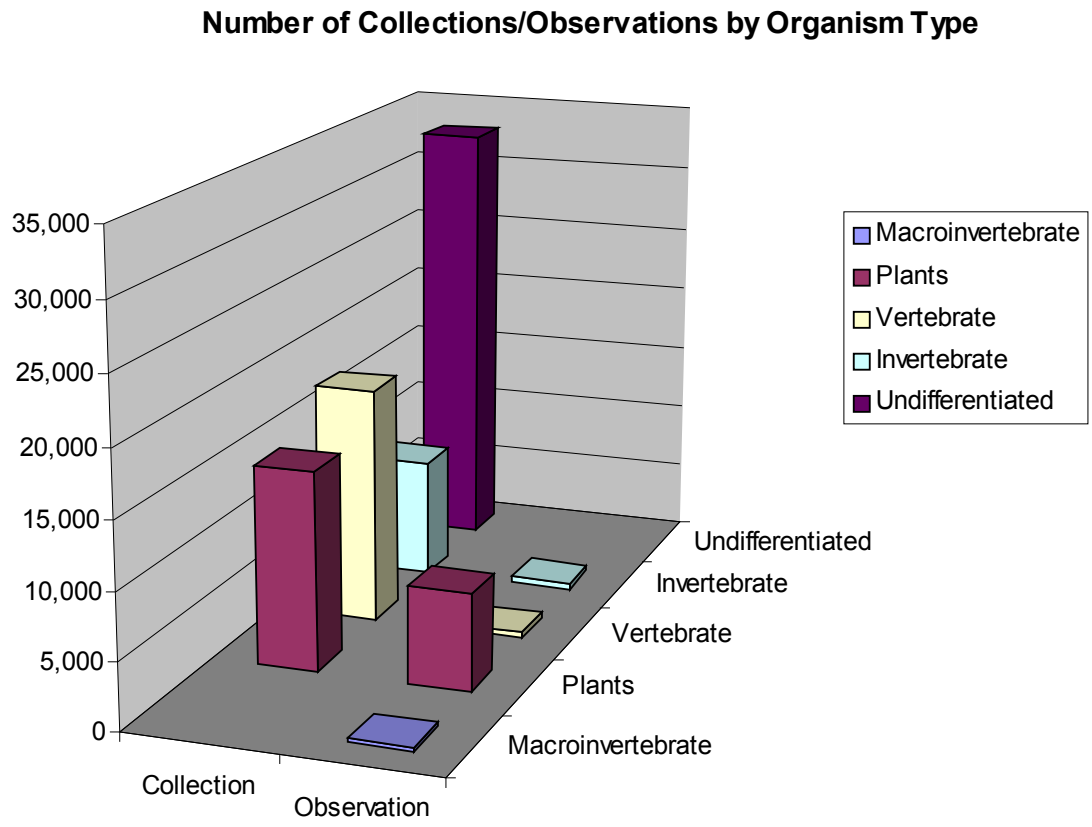
As mentioned above, numbers of taxa were also measured. The figures used were the number of taxa digitised (in the database) rather than the number of taxa in undigitised portions of collections/ observations. In total there were around 117,000 taxa represented. It is important to note that taxa were a much less accurate measure than number of collections/observations. In some instances these were 'ballpark' estimates. 34,000 taxa were from the Landcare Plant Names database and were not 'attached' to any collections/observations records. They were left out of the statistical analysis. For other taxa there was no way to avoid the risk of 'doubling up'. For example, a particular insect could easily be in the NZ Arthropod collection, as well as Te Papa and Otago's insect collections, its taxonomic name effectively being counted three times. For the purposes of detecting trends or patterns in the statistical analysis of survey responses, numbers of taxa were

simply added together. While not especially precise or 'scientific' this approach was still considered useful in understanding patterns in survey responses to questions on ease of export, data quality, and network connectivity for example.

Total numbers of taxa in the databases were:

	Plants	Vertebrate	Macroinvert.	Invertebrate	undifferentiated	Total
Collections	14,900	17,800		9,000	33,000	74,700
Observations	7,259	410	305	500		8,474
Total	**22,159**	**18,210**	**305**	**9,500**	**33,000**	**83,174**

This can be viewed as follows:



4.1.2 Percentage Digitised

In terms of percentages of collections/observations records actually in the databases these were as follows. As an issue orthogonal to the scope of this survey, analyses of these figures may give a view on the relative merits of connecting databases to GBIF as distinct from funding digitisation efforts.

Dataset Type	% digitised	% digitized (excl Invertebrates)
Collection	59%	79%
Observation	90%	89%

It should be noted that these figures are quite significantly skewed by a small number of datasets that have large volumes still undigitised. The very large majority of these data are invertebrate collection records, as can be seen in the differences in figures in the columns in the table above. To illustrate this, the following table represents proportion digitised when analysed by type of organism.

Organism Type	% digitised
Invertebrates	27%
Macroinvertebrates	80%
Plants	86%
Vertebrates	84%

4.2 Data Export

Survey respondents were given some background information on the GBIF exchange schemas, Darwin Core and ABCD (Access to Biological Collection Data).

They were asked questions to determine how easy it would be to export their bio diversity related data using GBIF exchange schema standards. The questions focused on:

a. How easy it would be to export just the bare minimum elements to make data useful in GBIF (taxon name, location, date, where held (if a specimen).

b. What other data was attached to each collection or observation record, and what proportion of non-mandatory fields could easily be exported to GBIF

The following ranking scheme was used in relation to mandatory field export:

Rank	Criteria
5	Our databases are already set up to export at least the GBIF mandatory fields in either Darwin Core or the ABCD exchange schema.
4	The bio data in this database is stored in such a way that it would be easy to export at least the mandatory fields in Darwin Core or ABCD exchange schemas.
3	GBIF mandatory fields in our bio data could be exported in Darwin Core or ABCD with a moderate amount of effort
2	Our bio data is in a format that would make export of GBIF mandatory fields in Darwin Core or ABCD a reasonable amount of effort.
1	It would be very difficult and time consuming to export even the mandatory fields from our bio data into Darwin Core or the ABCD schemas.

Most respondents were able to rank their ease of data export in terms of Darwin Core. Some were able to do so in terms of ABCD, however many were not given the much greater level of detail in ABCD. The following analysis is therefore based primarily on ease of export using Darwin Core.

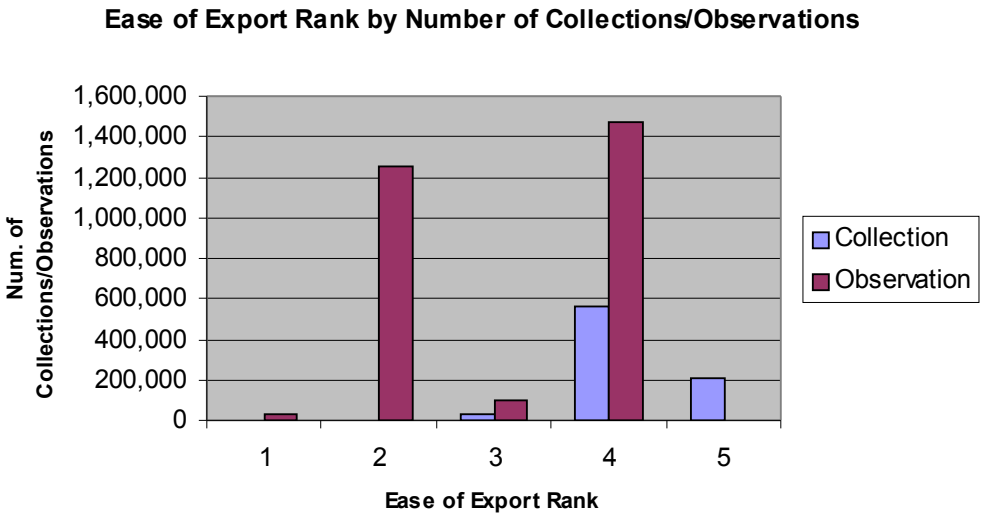
6% of databases ranked a 1 or a 2, 12% ranked as a 3, 79% ranked as a 4, and 3% ranked a 5.

Landcare Research, NIWA, Cawthron and Te Papa all rated their datasets as a 4. Otago Museum rated theirs as a 5. The Department of Conservation rated their datasets as a 3, with the exception of Bioweb Bird banding which was rated as a 2.

Another way of making sense of rankings is in terms of number of collections/observations and number of taxa. For both these numbers it is sensible to make a distinction between datasets that

are collections and those that are observations. This is because figures can be skewed considerably by some large volumes of observation data e.g. 1.25 million bird banding records, 1.26 million observations from vegetation plot survey data. There is also theoretically the possibility of skew in relation to arthropod records, as typically these collections number in the millions. The relatively low proportion of arthropod collections currently digitised however mitigates this potential skew at present.

Ease of mandatory field export rankings in terms of total number of collections/observations in the database are as follows:



This shows a large skew towards a 2 ranking for observations, however this is exclusively caused by the large number of observations in the DOC bird banding database. This anomaly disappears when ease of mandatory field export rank is considered in terms of number of taxa as shown below:

It could be concluded from all the above analyses that the large majority of data in New Zealand's major biodatabases could easily export GBIF mandatory fields.

Respondents were also asked for each primary collection or observation record, approximately (or exactly) how many data fields (i.e. fields per record) there were in total, and what proportion of these fields could easily be exported to GBIF. Numbers of fields, and proportion exportable to GBIF varied widely across different datasets. The total average number of additional fields that could be exported was 38. This broke down to 49 fields on average for collection datasets and 28 fields on average for observation datasets. When broken down by organisation type the average number of fields easily exported were 22 for DOC, 61 for Museums, 29 for Research Institutions and 16 for Universities. From this relatively informal analysis it can be concluded that the major biodatabases could relatively easily export quite a large number of non-mandatory fields to GBIF, and that Museum's datasets on average have a much larger number of potential fields exportable. This is not especially surprising as Darwin Core and ABCD were invented by, and on behalf of the museum community. This analysis of course says nothing about the relative scientific value of each field as it was not possible to go into this level of detail.

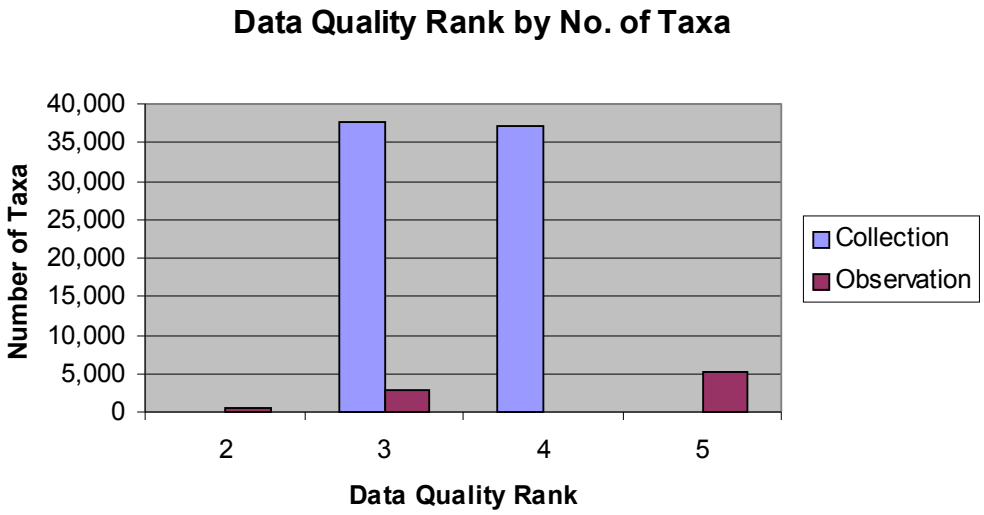
4.3 Accuracy/Quality Of Data

Survey respondents were asked to rank their databases in terms of data quality using the following table:

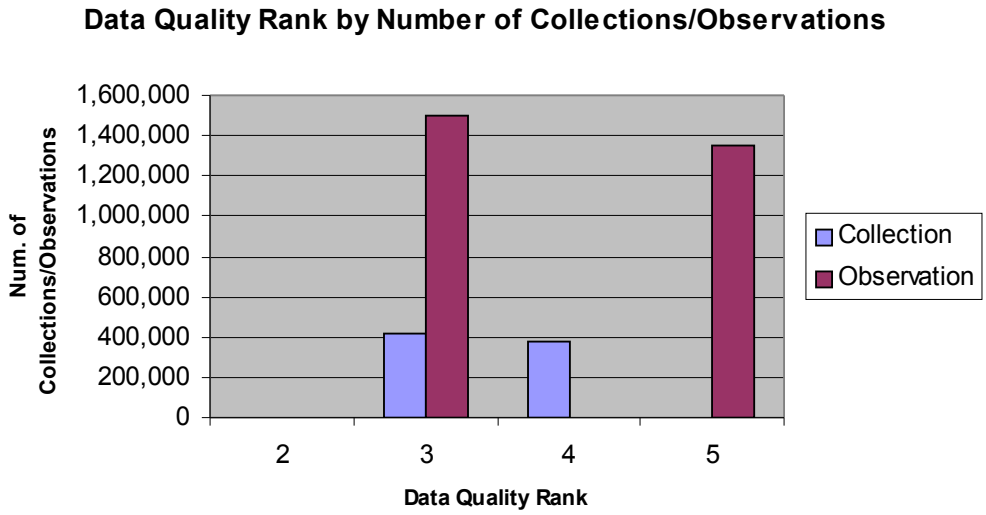
Rank	Criteria
5	Robust measures are in place to ensure quality of data. Data has defined certainty statements and data collection methods are subject to continuous monitoring to ensure standards. Data undergoes quality control procedures before it is introduced into the data management system. There is a rigorous process for ensuring the data retains integrity (complete and absent from introduced errors).
4	There are very good measures to ensure quality of this data. The data in this system has defined certainty statements and some measures are in place to monitor collection methods. Standard quality control measures exist for entry of data and these are enforced. A process exists to ensure data integrity however it is not rigorously enforced or monitored.
3	Reasonable measures are in place to ensure quality of data in the system. Data has defined certainty statements, however techniques to monitor collection processes are relatively adhoc. There are some basic quality control measures use when data is introduced to the system. There are no formal processes in place for ensuring data integrity, however we believe the data does not suffer from significant introduced errors.
2	There are rudimentary quality control measures for this system. Processes for collection of data are relatively standardised and coordinated, however there are no certainty statements in place for this data. No quality control measures exist for checking of data before it enters the system.
1	There are no quality management processes in place for this data. There is no measured level of certainty, and we are unable to guarantee that the same measures and instrument calibrations occurred for all the data in this system.

Responses are summarised as follows. 12% of databases ranked a 5, 15% a 4, 71% ranked a 3, and 3% ranked a 2. The existence of certainty statements are one of the key issues in combining research data or making it externally available. Given that 97% of databases were ranked as 3 or above this indicates that data quality may not be a significant barrier to integration of New Zealand's major biodatabases with GBIF. This however may be too much of a leap to make. It is important to note than in the surveyor's experience individual responses to questions of data quality can be quite variable. Often an 'inverse' effect applies where those who appear to have a very good understanding of data quality are likely to rank their datasets lower, while those who have less of an understanding tend to rank their datasets higher.

Considering data quality rankings in terms of collection/observation and taxa numbers the following can be seen:



The observations data ranked as 3 in this chart are all from Research Institute datasets (CRIs and Cawthron). The data ranked as 5 is from Department of Conservation datasets.



From this we can see that the majority of taxa are represented in databases that have a relatively high level of data quality.

It should be noted that the rankings given for data quality seem the most at risk of being subjective and may not, as they stand, be a useful measure, at least in terms of comparison across organisations. Without having analysed this thoroughly or scientifically there did seem to be a tendency for those organisations that understood data quality better to rate their datasets lower.

Comments in relation data quality were as follows:

- **Canterbury Museum - Zoology Collection:** We are currently at 3 for all data entered some areas of data are aiming to achieve standard 4 but are some way short of this milestone
- **Cawthron - Cawthron Macroinvertebrate Data:** 3 or 4 – given that the data is currently being reviewed. There are certainty statements in terms of depth of taxa identified to. There is an internal process for identification quality assurance.

- **Forest Research - Forest Research Herbarium:** Some of the data is quite old, and doesn't have grid references. Quality will improve with new database. For location data there is an indication of range (e.g. +/- 100m, 10k). There are no certainty statements in terms of skill/reputation of collector. Data is not double entered, often entered by the collector (relatively small herbarium). Data has been checked by mapping the locations spatially.
- **Landcare Research - 5 Minute Bird Counts Database:** Training mechanisms are in place, some quality control methods through individual trials balancing across a range of people to avoid bias.
- **Landcare Research - Allen Herbarium Specimen Database:** Data entered prior to July 2000 is at a poor level so would be more like a 2. We are certain that the data represents what's on the card. The rest would be a level 4. Everything is proofed and data entry is checked. There are DBA checks and informal integrity monitoring measures. There is a restricted set of users, and it is not possible to delete records.
- **Landcare Research - Mammal Distribution Database:** Uses a reliability rating for records. All existing range data is checked and rechecked by DOC and Regional Councils
- **Landcare Research - International Collection of Micro-organisms from Plants and Associated Databases (ICMP):** Anything with uncertainty is annotated.
- **Landcare Research - National Vegetation Survey Databank (NVS):** No vouchers, and the quality is variable. Quality depends on who did the observation, how difficult the taxa are, instances where the nomenclature are not up to date and are just as entered. Certainty is variable e.g. tagged trees have no defined certainty statements. Broad certainty statements could be made against specific taxa. For species names codes there is a look up table. There are various location range checks and altitude checks. There is no proofing or double entry. There are some logic checks and informal spot checks are done. There is no versioning or audit trail, although some paper records of changes on datasheets are kept.
- **Landcare Research - New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC):** Between 3 and 4. Done lots of checking of quality of the data post entry, e.g. sorting to see if wrong ones come out, plotting things, asking why. No defined certainty statements but checked through by specialists. Digitising isn't just a simple exercise of transcribing data, also requires lots of understanding (e.g. based on someone's handwriting, knowing where the collector had been, e.g. Rau, Ranui example). So double entry would be useful, but expert review is more important. A field that says 'how much can you trust this information' – e.g. an expert identification, reputable person, can't guarantee anything. Plan to do this for delivery of some information for TFBIS. Category B rather than category A for example. People generally need the information that goes around it. Getting QA done can work really well by making the data available and people externally checking it and giving feedback.
- **Landcare Research - New Zealand Fungal Herbarium and Associated Database:** There are unique fields that can't be doubled up accidentally for specimen number. A number of people enter the data so there is some possibility of error. There is no checking of identification of specimens when they are entered in. No certainty statements in terms of skill of determiner so some possibility for misinterpretation.
- **Landcare Research - Plant Names Database:** Data checking methods and error checking are used. There are business rules in the user interface that verify data. Bulk data entry is done in a 'holding pen' before being integrated into the main dataset.

- **Massey University - Dame Ella Campbell Herbarium (MPN) Database:** Mixture – standard data entry quality and integrity control, reasonable methods to ensure data quality; ad hoc collection monitoring.
- **Otago Museum - Natural Environment Collection:** Will be a 5 by the end of the year. Audit NZ and the Ministry of Culture and Heritage are coming to do a full audit of our collections and databases.
- **Te Papa - Natural Environment Collection:** Will implement a "confirmed by" field. Could assign certainty statements but it would be costly, some data sets quality control is good, some not. Changes to major fields such as 'Identification' are logged. There are no datasets that stand out as being much lower in quality than other data sets, but it would be fair to say that the mollusc data is of exceptionally high quality and the plant data of very high quality.

4.3.1 Data Dictionaries

For each database the following question was asked:

To what extent do you use Data Dictionaries as data content standards in relation to mandatory fields? Do you use for example:

- **Landcare Research Names Database [2](for taxonomic plant names)**
- **NZ Geographic Place Names Database[3] (for place names)**
- **Others…**

Four of the databases used external data dictionaries for species taxonomic names. This was done either by direct connection between the databases, or by batch transfer of a copy of the relevant data on an intermittent basis. For all these cases the data dictionary used was Landcare Research's Plant Names Database. It was noted that no authoritative national names databases seem to exist for invertebrates or macroinvertebrates.

Five of the databases derived their own internal data dictionaries for species taxonomic names from external published reference material such as NZ ornithological checklist, Journal of Herptofauna, CRI floras and faunas of NZ. These were databases held by DOC and Museums.

Six databases used their own internal data dictionaries for species taxonomic names. These were typically managed by taxonomic specialists in their fields, and were databases held by CRIs and Museums.

Three databases used external data dictionaries for NZ place names.

A table containing detailed findings from this question can be found in section 5.4.1 below.

4.4 Connectivity

It was explained to respondents that providing data to GBIF could involve anything from giving the New Zealand node manager (Landcare Research) an extract from their datasets once a year, through to setting up a 24x7 live connection to GBIF from their own networks.

Respondents were then asked questions about network connectivity and data connectivity.

4.4.1 Network Connectivity

GBIF is a distributed query network. In the future it is hoped that GBIF will be fully distributed and work in real time. Given current bandwidth and performance constraints GBIF currently uses a

[2] http://nzflora.landcareresearch.co.nz/

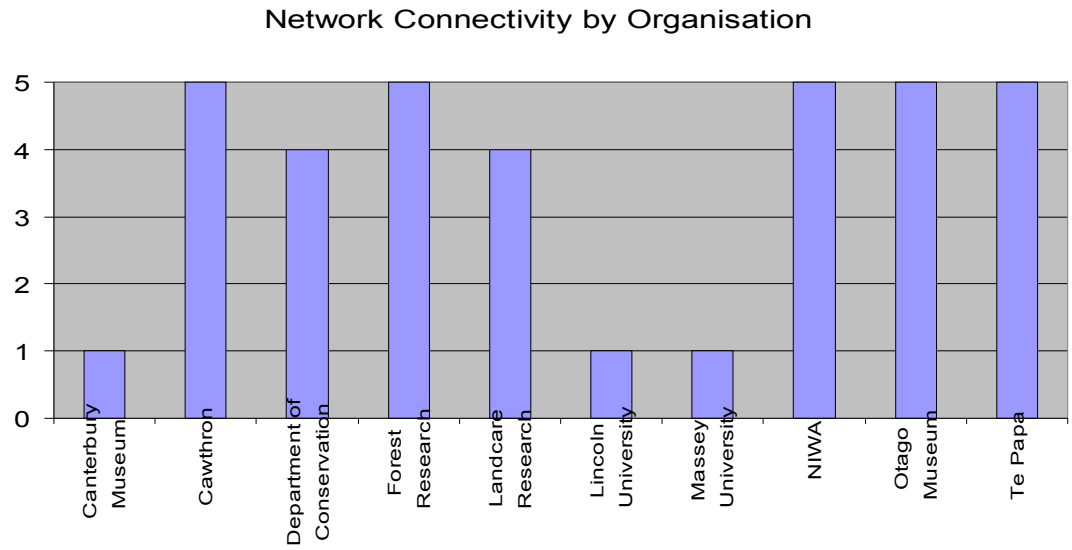
[3] http://www.linz.govt.nz/rcs/linz/pub/web/root/core/Placenames/SearchPlaceNames/downloaddataset/index.jsp?

harvester at the GBIF portal in Copenhagen which goes and gets the data and forms a central index. This provides a centralised cache of the high level data, then any drill down is done as a live query though the DiGIR or Biocase protocols into the original source datasets held at individual organisations.

Given this understanding respondents were asked to **rate their ability to connect their data to GBIF in terms of bandwidth and network connectivity using the following scale:**

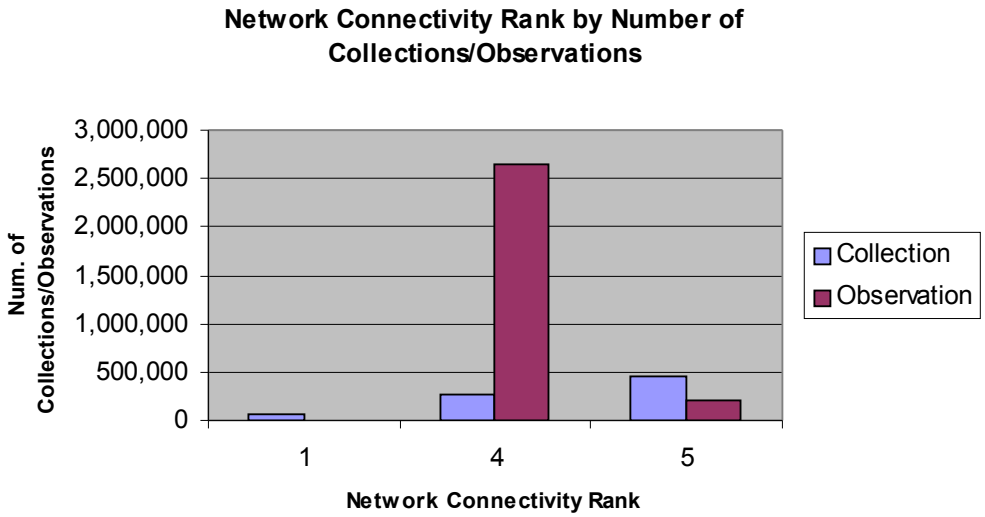
Rank	Criteria
5	Our network is available and supported for external access on a 24x7 basis. Available bandwidth is more than sufficient to handle the expected volume of queries to our data through the GBIF network.
4	Our network is available and supported for external access on a 9 to 5 basis. Available bandwidth is sufficient to handle the expected volume of queries to our data through the GBIF network.
3	Our network is available and supported for external access on a 9 to 5 basis. Available bandwidth should be sufficient to handle queries to our data through the GBIF network, however we are not able to give guarantees of response time.
2	Our network is available externally however bandwidth is limited and we would not be able to guarantee any level of response to direct queries to our data.
1	We do not have any ability to host GBIF connected data on our network.

Results were as follows:

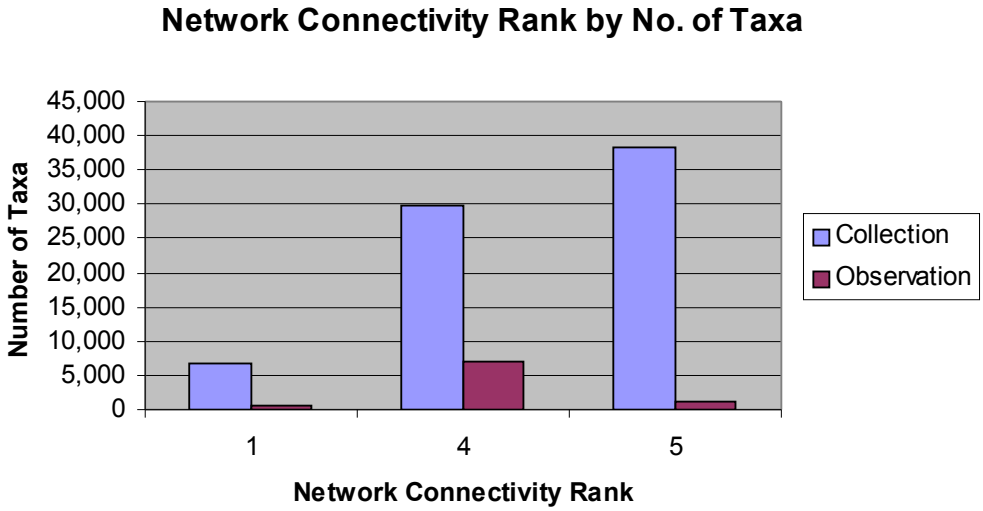


The majority of organisations reported they had network connectivity that would support a GBIF node. Three organisations reported a low potential level of network connectivity for their datasets. For Lincoln and Massey Universities this was due to the lack of current connectivity between the database and the network, and a lack of certainty about whether the University would provide this connectivity.

The following charts show network connectivity rankings in terms of number of collections/observations and number of taxa. This shows that the very large majority of collections/ observations and taxa are in organisations that could provide more than adequate network connectivity for a GBIF node.



and by taxa:



4.4.2 Data Connectivity

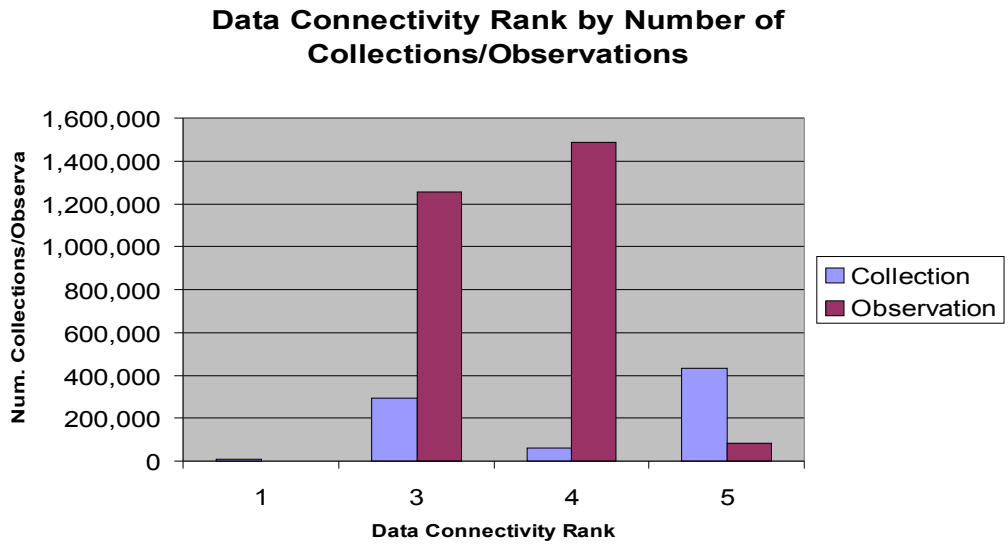
Participants were informed that for security reasons it is typical to put a GBIF data provider server outside the organisation's firewall, and on a server separate from the primary database servers. Based on this information respondents were asked to rate their ability to keep GBIF connected data up to date from their primary datasets using the following criteria:

Rank	Criteria
5	We are able to ensure that the data we expose to the GBIF network is updated on a daily basis from primary repositories in our organisation.
4	We are able to ensure that the data we expose to the GBIF network is updated on a monthly basis from primary repositories in our organisation.
3	We can host data ourselves, however we could only guarantee that it would be updated from our primary repositories a few times a year.
2	We could not host data ourselves, but we could provide data to be imported into a GBIF node (New Zealand or Copenhagen) three or four times a year.
1	We could not host data ourselves, but we could provide data to be imported into a GBIF node once a year.

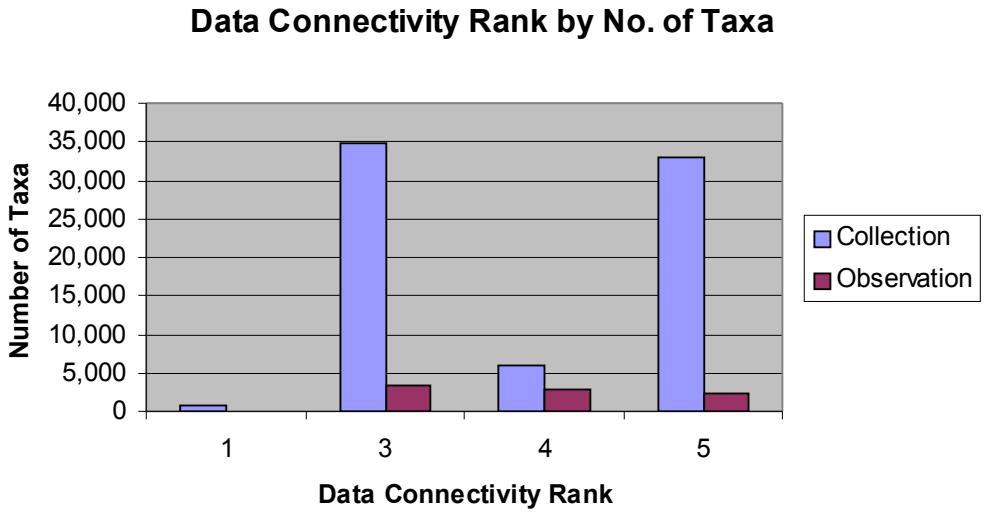
Quite coincidentally exactly the same number of datasets (11) ranked as 3, 4 and 5. Only one dataset (the Dame Ella Campbell Herbarium Database) ranked a 1 for data connectivity. When data connectivity rank was averaged across datasets held by each organisation Otago Museum and Te Papa ranked a 5, Cawthron and Niwa ranked a 4, DOC and Landcare Research ranked 3.5, and Canterbury Museum, Forest Research, and Lincoln University ranked a 3.

The high ranking from Otago Museum and Te Papa is almost certainly due to the fact that all their data was in one system (their Vernon collection management systems) rather than several different systems as is typical with CRIs, and that they already have plans/projects in place to make their collection management systems web accessible to the general public.

Data connectivity rank was also analysed in terms of number of collections/observations, and number of taxa. This shows that the great majority (>99%) of collections/observations surveyed were in organisations that could host data in their own GBIF connected node. 54% of collections (i.e. collection items) were in databases that could be updated from primary repositories to GBIF connected nodes on a daily basis. 34% of collections were in databases that could be updated to GBIF a few times a year. For observations, 45% were in systems that could be updated to GBIF a few times a year, 52% were in systems that could be updated to GBIF on a monthly basis, and only 3% were in systems that could be updated to GBIF daily. This is summarised in the following chart.



When data connectivity rank was considered in terms of number of taxa the following was seen.



To summarise, 47% of taxa were in collections that could be updated to GBIF a few times a year, 8% could be updated on a monthly basis, and 44% on a daily basis. Observations were more evenly distributed, with 39% updateable a few times a year, 32% monthly, and 28% daily. The taxa in collections that ranked 3 are contributed by the large numbers of taxa represented in the Landcare Research and Forest Research herbarium collections. Both these systems are currently stored in Access Databases. It should be noted that for both these systems there are plans to move the datasets to SQL-Server databases. It is likely this would increase the data connectivity rank to at least a 4 for these datasets.

4.5 Legal/Intellectual Property Constraints

Legal constraints and/or intellectual property rights can constrain an organisation's ability to share data through the GBIF network. For example, the data may have been collected with the permission of someone else (e.g. a landowner) under terms that do not permit sharing it outside the organisation, or a scientist may intend to publish and may not want the data available until after publication.

Respondents were asked to rank their datasets in terms of legal/intellectual property constraints according to the following criteria:

Rank	Criteria
5	The data we have that could potentially be useful to GBIF can be made accessible according to the conditions in the GBIF Data Sharing Agreement, there are no legal or IP constraints at all.
4	The large majority of the data we could put into GBIF can be made accessible domain according to the conditions in the GBIF Data Sharing Agreement, but there are some records or fields that we could not provide due to privacy, safety or IP reasons.
3	A reasonable proportion of the data we could put into GBIF is not encumbered by any legal or IP constraints. Some data that we currently have could be released within one or two years due to IP rights.
2	Of the data we have that could benefit GBIF, we could contribute a small subset, the rest we cannot due to privacy, safety or IP reasons.
1	We have primary collections and observation data, but we are unlikely to be able to put it into GBIF according to the conditions in the GBIF Data Sharing Agreement within the next 5 years because it is subject to legal, privacy or intellectual property constraints.

In total 1 database was ranked a 2, 14 databases were ranked 3, 16 ranked 4, and 3 ranked a 5.

When considered by organisation type the average score for Universities was 4.5, for Museums 3.9, and for Research Institutes 3.3. The comments made by respondents in relation to this question seem to indicate that the lower score from Research Institutes was due primarily to the fact they were holding data on other people's behalf (especially true for observations), or due to safety issues (biosecurity, rare or threatened plants). The issue of scientist's publishing rights and intellectual property while still important appeared secondary to the above issues.

Comments provided by respondents were:

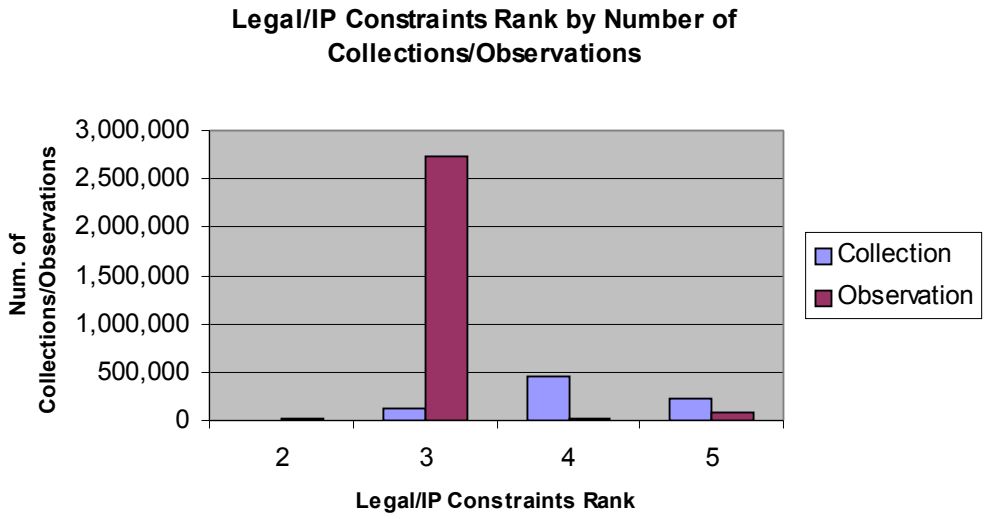
- **Canterbury Museum - Vertebrate Zoology Collection:** The phrasing of this question doesn't quite suit the Museum environment - Our major IP issue is that the museum gets acknowledged and that it is known that a specimen exists in the museum)
- **Cawthron - Cawthron Macroinvertebrate Data:** This is not in terms of IP issues, rather whether or not the data is confidential to the client. There is a field in the database that tracks whether it can be publicly available or whether it is client confidential. As a part of the contract we will be double-checking whether the historic data can be made available.
- **Forest Research - Forest Research Herbarium:** We would hide locations for threatened plant data, and would hide some data on plants on private property (e.g. valuable trees). We may just downgrade accuracy of location data on these records. There are no IP issues.
- **Landcare Research - 5 Minute Bird Counts Database:** Some counts done for private clients (AHB, Fletcher Forests, Electricity companies), so would have to clear with them. A lot is not

published so would want to have caveats. Very few would have no access, quite a few would require permission.

- **Landcare Research - Allen Herbarium Specimen Database:** Would have to restrict some collector fields for privacy reasons, and restrict locality for rare and threatened species
- **Landcare Research - International Collection of Micro-organisms from Plants and Associated Databases (ICMP):** Some data is deposited while the work is in progress, this would be suppressed until the work (including sometimes identification) is complete. Some of the data is used in work doing bio prospecting for companies, so this would also be suppressed. Apart from that ICMP was set up as an international repository so its purpose is to be freely available. The price for cultures is a service charge.
- **Landcare Research - National Vegetation Survey Databank (NVS):** Some would need permission in terms of IP or private land. At the moment there are only two options in our IP policy for NVS – 1. data that is freely available, 2. data for which you have to ask permission to get it entirely or any piece of it (i.e. to get into those datasets at all). We would like to introduce a new level for certain kinds of queries (i.e. for a limited subset that was not related to the kind of analysis the data collector wanted to do). For example Susan Wiser has collected data from banks peninsula for research on outcrops. Many species occurrence, environmental, and ecological factors were measured to ask scientific questions about how outcrops work ecologically. Someone contacted her who was doing research across 6 major tree species nationally. They wanted that data which was fine, as it was effectively an 'orthogonal slice of the data' in relation to the IP value/research purpose it was collected for. It would be good to have the freedom to allow those types of uses without having to go back and check every time with the collectors of the data.
- **Landcare Research - New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC):** 4 or 3. Current research projects won't be released until publication, information may be entered but there is uncertainty about the names until it's checked and revised. Don't think there are biosecurity/economic issues – e.g. MAF records aren't delivered on the Internet. But wouldn't release work in progress that might have these economic/biosecurity issues. Flag private/public. Some would be considered rare/endangered – location information isn't down to the precision that you can do with current GIS, only to closest km. Some minor IP issues from some researchers but only a few years, then normally public.
- **Landcare Research - New Zealand Fungal Herbarium and Associated Database:** We provide Crosby district, and map with a dot on it (fairly imprecise). We suppress locality descriptions for all records because they are often rare, dangerous or on private land. We suppress records of pathogens which may or may not occur in NZ (if there is uncertainty about whether they are here or not, e.g. early imprecise records, no strong evidence) for ones that would have an economic impact if revealed to be present in NZ.
- **Landcare Research - Plant Names Database:** Would need to restrict those not published yet or in a quality vetting process
- **Lincoln University - Centre of Research Excellence Database:** Could do all but aspects like gene sequence data for publishing reasons
- **Massey University - Dame Ella Campbell Herbarium (MPN) Database:** 5 is probably the answer, but I suspect that IP issues have never been addressed
- **NIWA - FBIS Data:** We would need to be sure we were not creating a commercial disadvantage

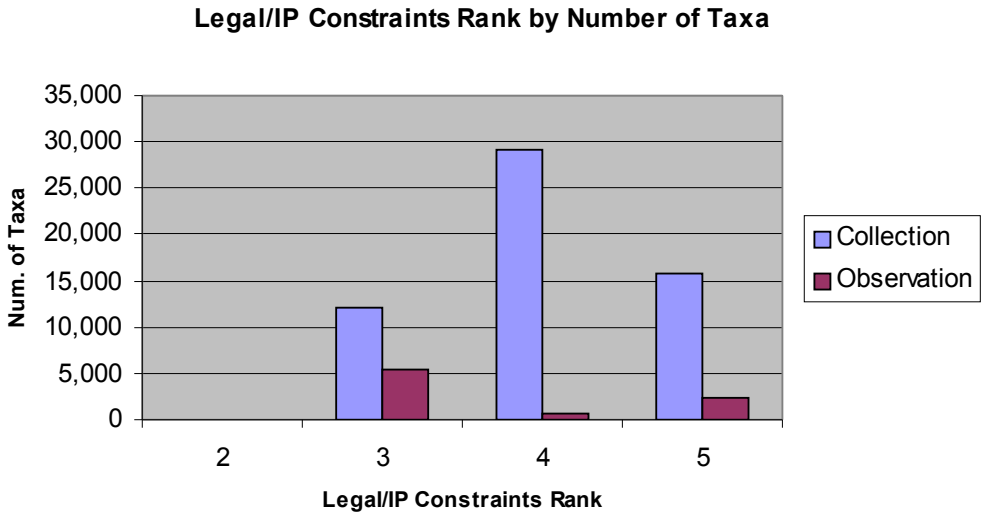
- **Otago Museum - Natural Environment Collection:** Minimal constraints anyway, and once something is put into the database the issues have been dealt with already.
- **Te Papa - Natural Environment Collection:** Some restrictions in terms of safety for rare and threatened species. Some restrictions for scientists IP for research currently underway.

When legal/IP constraints rank was considered in terms of number of collections/observations the following was seen:



95% of observations data was in systems that ranked a 3. This rank indicates "A reasonable proportion of the data we could put into GBIF is not encumbered by any legal or IP constraints". The reasons for this rank standing out seems due to observations often involving measurement on private land, and that observations data is often the subject of ongoing research.

The picture is different when considered by number of taxa:



For observations the majority are still ranked a 3, however a much greater proportion are ranked 4 or 5. For collections the majority are ranked 4 or 5.

When asked whether they had any specific policy regarding intellectual property or data use rights relating to their datasets participants responded as follows:

- **Canterbury Museum - Vertebrate Zoology Collection:** Yes but again the major issue is acknowledgement
- **Cawthron - Cawthron Macroinvertebrate Data:** Even if data is publicly available usually checked with the client.
- **Department of Conservation - Bioweb Bird banding:** Yes
- **Department of Conservation - Bioweb Threatened plants:** No
- **Landcare Research - National Vegetation Survey Databank (NVS):** Yes, this is formally stated on http://nvs.landcareresearch.co.nz/html/NVSprotocol.aspx. The protocol does not yet cover database to database access.
- **Landcare Research - New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC):** CRI Act, LCR policy, NSD policy document for all collections, nothing special for NZAC.

4.6 Willingness and Organisational Capacity

Respondent's attention was drawn to the fact that even if their data could be connected to GBIF in terms of exchange standards, data quality, IP issues and technical connectivity, there may be organisational or funding barriers. With this in mind they were asked to rate their organisation's willingness and capacity to connect data to GBIF as follows:

Rank	Criteria
5	We are very interested in exposing our data to the GBIF network. We are able to fund this ourselves both in terms of the initial set up, and on going support. There is high-level executive support for this that can reasonably be expected to continue.
4	We are interested in exposing our data to the GBIF network. One or two executives would actively support this. We are able to fund the majority of this ourselves but external funding would be of benefit.
3	We would like to expose or contribute our data to the GBIF network. Executives are aware of the benefits and reasons for doing this and are not opposed in any way. We may be able to fund this to a limited level ourselves, and would definitely be willing to do more if there was external funding available.
2	We would be willing to consider exposing or contributing our data to the GBIF network. Executives may consider this request but would be concerned about negative impacts on staff time and resources. We would not be able to fund this ourselves.
1	We are not interested in exposing our data to the GBIF network.

75% of datasets in total were in organisations that ranked their willingness and capacity as a 3, 18% as a 2, and 9% as a 4. When analysed organisation by organisation Canterbury Museum, Cawthron, Lincoln University and Massey University ranked a 2. DOC, Forest Research, Landcare Research, Niwa and Te Papa ranked a 3, and Otago Museum ranked a 4.

When considered in terms of number of collection/observations 95% of observations were in organisations ranked as a 3. For collections 54% ranked 3, and 36% ranked as a 4 in terms of willingness. When considered in terms of number of taxa the ratios were similar with 91% in organisations ranked as a 3, and for collections with 19% ranked as 2, 61% ranked 3, and 31% ranked 4.

Comments received indicate that willingness and capacity is almost purely a funding question. Lower levels of willingness/capacity for Universities may be due to the relatively lower level of importance or profile given to research datasets as compared to CRIs or Museums. Universities

also seem to have a much lower level of biodiversity informatics infrastructure than CRIs or Museums.

Comments received in relation to this question are as follows

- **Cawthron - Cawthron Macroinvertebrate Data:** Low willingness to connect ourselves however these data will all be made available through FBIS
- **Forest Research - Forest Research Herbarium:** Would like to connect to GBIF but haven't done it yet. We might be able to fund with the FRST funding we have at the moment. Executives wouldn't actively champion but wouldn't oppose it at all, some funding might help or make it happen sooner.
- **Landcare Research - 5 Minute Bird Counts Database:** Landcare Research would probably not be prepared to fund it, would have to be totally funded or majority funded to do this.
- **Landcare Research - Allen Herbarium Specimen Database:** There is always a trade off between making data available, and the risks of missaplication of the data. GBIF could be a good discovery mechanism to encourage/require users to talk to Landcare Research to get more complete datasets and understand the caveats
- **Landcare Research:**
    - **Mammal Distribution Database:** Will quite likely sit under DOC's NHMS
    - **International Collection of Micro-organisms from Plants and Associated Databases (ICMP):** external funding is the big thing here
    - **New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC):** external funding is the big thing here
- **Massey University - Dame Ella Campbell Herbarium (MPN) Database:** Low priority compared with databasing our existing collection
- **NIWA - FBIS Data:** NIWA would be somewhere between 3-4
- **Otago Museum - Natural Environment Collection**: On the high end of willingness, but external funding would help. We are likely to do it eventually anyway, but it will take longer without external funding.
- **Te Papa - Natural Environment Collection**: - Willingness has not been tested yet. There are implications in that more collection loan requests would be likely and there are associated cost implications. We only have a small IT team. We would have to have initial internal discussions, however most people would want to do this including executive management.

5 Appendix

This appendix contains references used in the preparation of the report, a set of statistical data on regional council funding, and detailed findings from this survey.

Detailed findings include a list of data dictionaries used by individual datasets, along with full information on each dataset and each organisation's responses to more general questions as collected during the survey.

5.1 Further References

New Zealand Biodiversity Strategy - http://www.biodiversity.govt.nz/

TFBIS - http://www.biodiversity.govt.nz/land/nzbs/tfbis/tfbis/index.html

GBIF - http://www.gbif.org

New Zealand GBIF node - http://www.gbif.org.nz

Darwin Core - http://darwincore.calacademy.org/

ABCD standard - http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/Schema/default.htm

Other TDWG standards - http://www.tdwg.org/subgrops.html

References used to find the databases:

A Nationally Significant Databases and Collections Providers' Group - http://natsigdc.landcareresearch.co.nz/

New Zealand National Herbarium Network - http://www.nzherbaria.org.nz/

Statistics NZ Directory of Environmental Databases - http://www2.stats.govt.nz/domino/external/web/catv2.nsf/byOrganisation?openview

5.2 Glossary

Research Data – defined as in the U.S. National Institutes of Health definition of final research data: "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings". It should be noted that in this paper the term 'data' is used in the scientific, plural sense, i.e. this 'datum', these 'data'.

Biodata – research data relating to biology and/or biodiversity, often relating to collections or observations of plants, mammals and invertebrates.

Biodatabase – a database containing biodata

Geographic Information System (GIS) – a computerized system for combining, displaying, and analyzing geographic data. GIS produces maps for environmental planning and management by integrating physical and biological information (soils, vegetation, hydrology, living resources, and so forth) and cultural information (population, political boundaries, roads, bank and shoreline development, etc)

Collections – a collection is a set of specimens collected in the field and held in a particular institution. For some collections, individual collection items are 'lots' which contain more than one specimen. Each collection item is typically labeled in some way. In well-managed collections collection items are independently identified, vouchered and metadata stored in a collection management database. Collections are often held in climate controlled environments. For this report a dataset can be a collection dataset (i.e. it contains records of many individual collection items). In this report phrases like 'number of collections' is shorthand for number of collection items, i.e. 5 collections is synonymous with 5 specimens.

Observations – an observation is a record of a sighting of a particular organism in the field. In this report some datasets are referred to as observation datasets (i.e. containing records of many observations).

Taxonomic names – the scientific name of a plant or organism, including its place within the Linnaean hierarchy (Kingdom – Phylum – Class – Order – Family – Genus – Species), and often including vernacular names and synonyms

Taxa – (singular = taxon) the named classification unit to which individuals or sets of species are assigned, such as species, genus and order.

Darwin Core – a simple set of data element definitions designed to support the sharing and integration of primary biodiversity data.

ABCD – the Access to Biological Collections Data (ABCD) Schema is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data).

Data Dictionary – a controlled list of items that restrict entry of data into a particular field in a database to only the items in that list.

Informatics – A field of study that focuses on the use of technology for improving access to and utilization of information. Health informatics for example is the systematic study of information in the healthcare delivery system, how it is captured, retrieved, and used in making decisions, as well as the tools and methods used to manage this information and support decisions.

Middleware – Software that sits between two or more types of software and translates information between them. Middleware can cover a broad spectrum of software and generally sits between an application and an operating system, a network operating system, or a database management system. In the context of research data middleware applications often include tools that enable indexing, archiving, discovery, analysis, integration, management and preservation of large heterogeneous distributed data repositories.

Web services – Web services let computers talk to one another over the Internet, allowing computer programs to exchange information by eliminating barriers such as different hardware platforms, software languages, and operating systems that usually make different programs incompatible. Web services make it easier to share information, data, and services, as well as making it cheaper and easier for organisations to work with on-line partners.

5.3 Datasets Noted But Not Surveyed In Depth

The following datasets were noted during the survey process but were not analysed in any depth:

Organisation and Database name	Description	Reason for not analysing fully
Landcare Research - Nematode collection	1000 specimens, quite a number of which are not from New Zealand. Currently just one flat file, would require lots of work on it to turn it into a Relational Database.	Too small, not currently in a relational database.
Landcare Research Nga Tipu Whakaoranga - People Plants Infobase	An infobase on the traditional uses of New Zealand native plants by Maori.	Almost everything in this database is covered in the Allen Herbarium Database
DOC Other databases	5 minute bird counts; bats, kakapo, kiwi datasets for specific studies; kiwi database;	Too small, not in a full relational

	whale strandings; whale sightings; Pestlink (rats, possums, goats)	database, or not enough time to analyse.
Lincoln University Entomology Research Museum Database John Marris	Database for the Entomology Research Museum that is in its infancy, with only a couple of thousand insect specimen records to date. This data, which has mainly been input courtesy of TFBIS funding, is presently on an excel spreadsheet but we have been trialing Biota (BiotaApp2) database software and, at this stage, intend continuing with this.	Too small, not in a full relational database.
University of Canterbury Insect Survey Database Raphael Didham	Hope River Forest Fragmentation Invertebrate Database, Hurunui District. Beetle data, site description data, lepidoptera. spiders, range of families from within diptera, thysenoptera, some other families half completed. Conducted from Nov 2000 to March 2001 over 3 valleys 400sqkm, using flight intercept trap (ground and forest canopy) sampling and soil litter sampling. Ground beetles 35,400 individual records, 893 species, 700,000 sorted invertebrates (not in the database yet), Other data includes wood decomposition, crude vegetation analysis, microclimate measurements (temperature, vapour pressure deficit, relative light intensity),	Too small (to be considered a major biodatabase at least)

5.4 Detailed Findings

This section contains a record of survey response data for data dictionaries, and following that tables containing metadata, and all responses for each individual database.

5.4.1 Data Dictionaries

- **Canterbury Museum - Vertebrate Zoology Collection**
  Data Dictionaries Used: NZ Geographic Places database and various specific taxonomic databases (mostly in the ITIS framework)
- **Cawthron - Cawthron Macroinvertebrate Data**
  Data Dictionaries Used: Internal tables for taxa, sample method, analysis method
  Notes: Internal taxa table as 'master list', any names changes follow through on that list. When this was developed we looked at international databases for data structure standards for taxonomy but didn't find anything consistent (ITIS, taxonomic serial number for species table, Species 2000). We came up with one that sorts everything phylogenetically and has a standardised format so you can tell something about the critter by the ID. This is consistent across all taxonomic databases at Cawthron. There are lookup tables for sample method, analysis method. We are using the full 6 digit NZ mapgrid so there is no need for place names. We will change to NZTM in due course.
- **Department of Conservation - Bioweb Bird banding**
  Data Dictionaries Used: Uses NZ ornithological checklist

- **Department of Conservation - Bioweb Herpetofauna**
  Data Dictionaries Used: Uses Journal of Herptofauna, and specialist panel convened by DOC
- **Department of Conservation - Bioweb Threatened plants**
  Data Dictionaries Used: Uses Landcare standards for names, and a specialist panel convened by DOC for threat status.
- **Department of Conservation - Bioweb Weeds**
  Data Dictionaries Used: Uses published botanical references
- **Forest Research - Forest Research Herbarium**
  Data Dictionaries Used: Ecological districts (internal, read only), plant names family level (internal, editable by some users), authorities (internal, can be edited/added to)
  Notes: Landcare Research's Names DB could be used, but the Forest Research Herbarium database deals with a lot of cultivated plants that wouldn't be in LCR Plant names, so this might be tricky. We are planning to migrate the system to on SQL-Server which will make our names data more widely accessible internally. It will be interesting to see how people make use of this within Forest Research.
- **Landcare Research - 5 Minute Bird Counts Database**
  Data Dictionaries Used: None
  Notes: May be included in new database
- **Landcare Research - Allen Herbarium Specimen Database**
  Data Dictionaries Used: Names, Countries, Land districts, Ecological districts, Ecological keywords
  Notes: Names has an update dictionary alongside a verbatim field
- **Landcare Research - International Collection of Micro-organisms from Plants and Associated Databases (ICMP)**
  Data Dictionaries Used: Names module for fungi and bacteria
  Notes: Names module for fungi and bacteria. If a person can't find a match there is a process for entry of new names into the names module. This includes recording the citation and authors. The Names module will be integrated in with the plant names system as a move to a more sophisticated platform in the future. Only broad localities are recorded so place names are not important.
- **Landcare Research - Mammal Distribution Database**
  Data Dictionaries Used: None
  Notes: Manual checking against topo maps
- **Landcare Research - National Vegetation Survey Databank (NVS)**
  Data Dictionaries Used: Landcare Research Plant Names Database, some internal data dictionaries
- **Landcare Research - New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC)**
  Data Dictionaries Used: NZ Gazeteer
  Notes: For geographic names use NZ gazeteer. There is a names module, based on what was done for Species 2000. Have a separate field for recording verbatim/as it is, but primary interpretation is from a standard list. The set of names will be provided through NZBugs portal so other can use it (similar to plant names).
- **Landcare Research - New Zealand Fungal Herbarium and Associated Database**
  Data Dictionaries Used:
  Notes: Landcare Plant Names database is not used 'live' for host names, this data dictionary is maintained within the databases. If we come across a host name that's not there we just add it.

We are going through the process of hooking directly to the Plant Names Database at the moment. Fungi names are from the Names module for this system. Place names are normally locality, this is enterpreted using topomap and grid coordinates entered. There is no indication of scale of accuracy for this in the database.

- **Landcare Research - Plant Names Database**
  Data Dictionaries Used: Country, Region, Key word, Association, Biostatus categories
  Notes: All are static and internal
- **Lincoln University - Centre of Research Excellence Database**
  Data Dictionaries Used:
  Notes: Will be hooked into a Landcare Research database for invertebrate names. No place names are recorded.
- **Massey University - Dame Ella Campbell Herbarium (MPN) Database**
  Data Dictionaries Used: Landcare names database
  Notes: Use Landcare names database, and misc. others for checking taxonomy
- **Otago Museum - Natural Environment Collection**
  Data Dictionaries Used: Internal
  Notes: We use the CRI floras and faunas of NZ, and gazetteers for place names. All are internal copies based on CRI lists, journals, publications and web sites. We do systematics research and do change names and update them. We try to make sure these name changes are fed back to other organisations holding lists of taxonomic names.
- **Te Papa - Natural Environment Collection**
  Data Dictionaries Used: Names (internal), LINZ, Getty Institute
  Notes: Internal hierarchies of names as data dictionary. Place names – will use LINZ and Getty Institute thesaurus for international names. Use NZMS260 mapping sheets. Databases of collectors used as a data dictionary. Could use Landcare Research's list of plant names but only in batch delivered mode.

5.4.2 Cawthron - Cawthron Macroinvertebrate Data

Title	Cawthron Macroinvertebrate Data
Abstract	Cawthron has a large amount of macroinvertebrate data, primarily counts and related information from samples. There is a TFBIS project undersay to update all historical data into a new system. This is expected to be finished within the 2005 calendar year. This is involves introduction of the data into Cawthron's core systems, and a once off transfer of those data to NIWA's FBIS.
Contact Person	Paul Barter
Organisation	Cawthron
Temporal Extent	Mid 1980s to present. The temporal extent of the data we're going to enter into FBIS depends a bit on data quality (i.e. some of the earlier work may not be of suitable quality for entry into FBIS).
Spatial Extent	Predominantly South Island, New Zealand, concentrated on the top of the South Island

Volume	Total Number of Collections or Observations – 100,000 Number of Collections or Observations in Database – 100,000 Number of Taxa in Database – 200
Ability to export of the mandatory fields required by GBIF	Rank: 4 Notes: Under the current contract we are set up to do import of historical data and one flat file export, into FBIS. To do this we are importing it all into an existing database structure used for internal data management and reporting, then doing a one off export into FBIS.
Number of data fields collection or observation record	30 Notes: Includes sampling methods (sampler type, units, length, volume etc), client information (who collected for), project information, whether its publically available, site information (location, date, time, depth, units), collector, notes, analysis methods (count data vs point counts etc), species taxonomic information (linaen & subclass, subfamily), reproduction and feeding (for some species), datatype (summed, counts by area, unit qualifiers), MCI, QMCI, Rapid Assessment Method, Cawthron contact, Cawthron report number.
Proportion of fields per record that could easily be exported to GBIF	90% Notes: Estimate 90% would fit with Darwin core or could easily be converted. The 10% is mostly data that would not be of interest to GBIF.
Data Quality	Rank: 3 Notes: 3 or 4 – given that the data is currently being reviewed. There are certainty statements in terms of depth of taxa identified to. There is an internal process for identification quality assurance.
Data Dictionaries	Dictionaries used: Internal tables for taxa, sample method, analysis method Notes: Internal taxa table as 'master list', any names changes follow through on that list. When this was developed we looked at international databases for data structure standards for taxonomy but didn't find anything consistent (ITIS, taxonomic serial number for species table, Species 2000). We came up with one that sorts everything phylogenetically and has a standardised format so you can tell something about the critter by the ID. This is consistent across all taxonomic databases at Cawthron. There are lookup tables for sample method, analysis method. We are using the full 6 digit NZ mapgrid so there is no need for place names. We will change to NZTM in due course.
Connectivity	Network Connectivity Rank: 5 Notes: Once data is in FBIS Data Connectivity Rank: 4 Notes: Future transfers of information would be from our internal database into FBIS (we will have the structures in place to initiate future transfers between the two databases by the end of our TFBIS project). How often we transfer public data will depend on the cost of doing the transfer (which we should have a feel for by the end of our TFBIS project).

Legal/IP Constraints	Rank: 3 Notes: This is not in terms of IP issues, rather whether or not the data is confidential to the client. There is a field in the database that tracks whether it can be publicly available or whether it is client confidential. As a part of the contract we will be double-checking whether the historic data can be made available. IP Policy: Even if data is publicly available usually checked with the client.
Organisational Willingness	Rank: 2 Notes: There would be a fairly low willingness to connect this to GBIF ourselves, however the data will be made available through FBIS and could be connected to GBIF that way.
Other Notes:	

5.4.3 Forest Research - Forest Research Herbarium

Title	Forest Research Herbarium
Abstract	Founded in 1946, the herbarium contains approximately 25,000 specimens. It includes native, adventive and cultivated plants found in New Zealand, especially vascular plants of native and exotic forests, as well as specimens of forest trees from Australia, Fiji, Samoa, Vanuatu and Mexico. There are extensive collections of conifers, eucalypts and other introduced plantation tree species. Information from all 25,000 specimens is held on a computer database. External specimens are used for tree/plant identification, vouchering. Primary use is tree identification for the forest industry. It is also used for systematics, and used by the forest health group. Specimens are also collected from around the ports in terms of pests identification.
Contact Person	Chris Ecroyd
Organisation	Forest Research
Temporal Extent	1946 to present
Spatial Extent	Primarily New Zealand, some overseas specimens used for comparison and pest identification purposes.
Volume	Total Number of Collections or Observations – 25,000 Number of Collections or Observations in Database – 25,000 Number of Taxa in Database - 0
Ability to export the mandatory fields required by GBIF	Rank: 3 Notes: Database is currently being converted to an Access DB (was in Advanced Revelation), next step is to convert to SQL Server. Currently using HISPID (Herbarium Information Standards and Protocols for Interchange of Data), which derives from Darwin Core.
Number of data fields in collection or observation record	50 Notes: e.g. collectors name, collectors date, altitude, map reference, location (by ecological districts) etc
Proportion of fields per record that could easily be exported to GBIF	60% Notes: About 30 could be easily be exported (measured by comparison). Until we've tried it we wouldn't know exactly.

Data Quality	Rank: 4 Notes: Some of the data is quite old, and doesn't have grid references. Quality will improve with new database. For location data there is an indication of range (e.g. +/- 100m, 10k). There are no certainty statements in terms of skill/reputation of collector. Data is not double entered, often entered by the collector (relatively small herbarium). Data has been checked by mapping the locations spatially.
Data Dictionaries	Dictionaries used: Ecological districts (internal, read only), plant names family level (internal, editable by some users), authorities (internal, can be edited/added to) Notes: Landcare Research's Names DB could be used, but the Forest Research Herbarium database deals with a lot of cultivated plants that wouldn't be in LCR Plant names, so this might be tricky. We are planning to migrate the system to on SQL-Server which will make our names data more widely accessible internally. It will be interesting to see how people make use of this within Forest Research.
Connectivity	Network Connectivity Rank: 5 Notes: Once on to a server database would see it as a 5. Data Connectivity Rank: 3 Notes: Might be a 3 or a 4.
Legal/IP Constraints	Rank: 4 Notes: We would hide locations for threatened plant data, and would hide some data on plants on private property (e.g. valuable trees). We may just downgrade accuracy of location data on these records. There are no IP issues.
Organisational Willingness	Rank: 3 Notes: Would like to connect to GBIF but haven't done it yet. We might be able to fund with the FRST funding we have at the moment. Executives wouldn't actively champion but wouldn't oppose it at all, some funding might help or make it happen sooner.

5.4.4 Landcare Research - Allen Herbarium Specimen Database

Title	Allen Herbarium Specimen Database
Abstract	The Specimen Database is used to store and retrieve herbarium specimen information and to generate specimen labels. It is the largest database at the herbarium, currently containing over 115,000 records, or approximately 20% of the specimens in the herbarium. Specimens are added to the Database according to research and conservation priorities.
Contact Person	Aaron Wilton
Organisation	Landcare Research
Temporal Extent	16th Century to present
Spatial Extent	New Zealand, and exotic plants with geographic or systematic links
Volume	Total Number of Collections or Observations – 550,000 Number of Collections or Observations in Database – 130,000 Number of Taxa in Database – 15,000

Ability to export the mandatory fields required by GBIF	Rank: 4 Notes: Currently set up but is a manual process to export to Darwin Core. For ABCD it would be a 3.
Number of data fields in collection or observation record	50 Notes: 15 – 20 fields on abundance habitat notes, duplicates, taxonomic info, locality, info, altitude, map references, host species, determiner (who identified the specimen), when identified, notes, associated species (growing nearby), determination history
Proportion of fields per record that could easily be exported to GBIF	40% Notes: Could export the majority of these easily. There could be some challenges with name format for collector
Data Quality	Rank: 4 Notes: Data entered prior to July 2000 is at a poor level so would be more like a 2. We are certain that the data represents what's on the card. The rest would be a level 4. Everything is proofed and data entry is checked. There are DBA checks and informal integrity monitoring measures. There is a restricted set of users, and it is not possible to delete records.
Data Dictionaries	Dictionaries used: Names, Countries, Land districts, Ecological districts, Ecological keywords Notes: Names has an update dictionary alongside a verbatim field
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 3 Notes: In MS Access at present
Legal/IP Constraints	Rank: 4 Notes: Would have to restrict some collector fields for privacy reasons, and restrict locality for rare and threatened species IP Policy: CRI Act
Organisational Willingness	Rank: 3 Notes: There is always a trade off between making data available, and the risks of misapplication of the data. GBIF could be a good discovery mechanism to encourage/require users to talk to Landcare Research to get more complete datasets and understand the caveats
Other Notes:	

5.4.5 Landcare Research - Plant Names Database

Title	Plant Names Database
Abstract	The Plant Names Database records the scientific names of plant taxa in the New Zealand flora. It includes the current names for lichens, liverworts, mosses, ferns, and seed plants that are wild in New Zealand. The database also includes the synonyms for the New Zealand mosses, lichens and ferns.
Contact Person	Aaron Wilton
Organisation	Landcare Research
Temporal Extent	
Spatial Extent	New Zealand

Volume	Total Number of Collections or Observations - 0 Number of Collections or Observations in Database - 0 Number of Taxa in Database - 34,000
Ability to export the mandatory fields required by GBIF	Rank: 4 Notes:
Number of data fields in collection or observation record	0 Notes: 20 additional nomenclateral and taxonomic related fields (e.g. current name, is it valid, when was it published, where. These are in Linnaean Core, we will do all others in ABCD.
Proportion of fields per record that could easily be exported to GBIF	0 Notes:
Data Quality	Rank: 3 Notes: Data checking methods and error checking are used. There are business rules in the user interface that verify data. Bulk data entry is done in a 'holding pen' before being integrated into the main dataset.
Data Dictionaries	Dictionaries used: Country, Region, Key word, Association, Biostatus categories Notes: All are static and internal
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 5 Notes: Could update weekly
Legal/IP Constraints	Rank: 4 Notes: Would need to restrict those not published yet or in a quality vetting process IP Policy: Data agreement with Te Papa & Canterbury University
Organisational Willingness	Rank: 3
Other Notes:	

5.4.6 Landcare Research - New Zealand Fungal Herbarium and Associated Database

Title	New Zealand Fungal Herbarium and Associated Database

Abstract	In 2001 a suite of legacy databases relating to mycology, bacteriology, systematics and pathology were consolidated into an integrated information resource and some (but not all) of these data, were made available through the NZFUNGI website. The development of NZFUNGI was one output from the Database Integration Project (DIP) which ran from 1999-2004 and was partly funded by the Foundation for Research Science and Technology. DIP was focussed on redevelopment and accessibility of the data contained in the seven FRST Nationally Significant Databases & Collections maintained by Landcare Research. The National Fungal Herbarium (PDD) and the International Collection of Micro-organisms from Plants (ICMP) are two of those collections/databases, and the data from the dried and living collections of fungi and bacteria form one part of the NZFUNGI information resource. The database currently contains six related categories of information: names, literature, herbarium collections, cultures, hosts and lists.
Contact Person	Peter Johnston
Organisation	Landcare Research
Temporal Extent	mid 1800s to present
Spatial Extent	Approximately 66% of the collection is New Zealand, around 17% is from the Pacific Islands (this is the primary repository for them), the remaining 17% is from the rest of the world (exchange specimens, and for comparison to NZ specimen purposes).
Volume	Total Number of Collections or Observations – 70,000 Number of Collections or Observations in Database – 70,000 Number of Taxa in Database – 6,000-7,000
Ability to export the mandatory fields required by GBIF	Rank: 4 Notes:
Number of data fields in collection or observation record	80 Notes: 80 fields approximately including: identifier, date identified, date collected, where it was collected, the kind of habitat collected in, the host it was associated with, secondary associations, Crosby district, type status, loan information (who borrowed, when due for return)
Proportion of fields per record that could easily be exported to GBIF	80% Notes: We have quite a number of fields for associated organisms (a thing growing on a thing, growing on a thing). This is not represented in Darwin Core. There are also a number of fields related to collection management that wouldn't be of interest to GBIF. Geospatial standards could be simplified down from 15 to: lat/long/ or gridref, datum or mapsheet, accuracy and precision.
Data Quality	Rank: 3 Notes: There are unique fields that can't be doubled up accidentally for specimen number. A number of people enter the data so there is some possibility of error. There is no checking of identification of specimens when they are entered in. No certainty statements in terms of skill of determiner so some possibility for misinterpretation.

Data Dictionaries	Dictionaries used: internal Names module Notes: Landcare Plant Names database is not used 'live' for host names, this data dictionary is maintained within the databases. If we come across a host name that's not there we just add it. We are going through the process of hooking directly to the Plant Names Database at the moment. Fungi names are from the Names module for this system. Place names are normally locality, this is enterpreted using topomap and grid coordinates entered. There is no indication of scale of accuracy for this in the database.
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 3 Notes: The database is currently in stored in Microsoft Access. Ideally it should be converted to SQL Server. The copy on the web site gets updated every couple of months. Descriptions data – don't know how compatible this is with GBIF? Text fields that link to names.
Legal/IP Constraints	Rank: 4 Notes: We provide Crosby district, and map with a dot on it (fairly imprecise). We suppress locality descriptions for all records because they are often rare, dangerous or on private land. We suppress records of pathogens which may or may not occur in NZ (if there is uncertainty about whether they are here or not, e.g. early imprecise records, no strong evidence) for ones that would have an economic impact if revealed to be present in NZ. IP Policy: Landcare Research generic policy on IP, nothing specific for this dataset
Organisational Willingness	Rank: 3
Other Notes:	

5.4.7 Landcare Research - International Collection of Micro-organisms from Plants and Associated Databases (ICMP)

Title	International Collection of Micro-organisms from Plants and Associated Databases (ICMP)
Abstract	ICMP is a major international collection of plant bacteria, and a repository for micro-organisms of plant and animal origin of the New Zealand Reference Culture Collection. It also includes cultures of the world's bacterial and fungal plant pathogens and of other micro-organisms closely associated with plants. It contains over 12,000 strains of micro-organisms.
Contact Person	Shaun Pennycook
Organisation	Landcare Research
Temporal Extent	1960's to present.
Spatial Extent	New Zealand, Global
Volume	Total Number of Collections or Observations – 12,000 Number of Collections or Observations in Database – 12,000 Number of Taxa in Database - 1,800
Ability to export the mandatory fields required by GBIF	Rank: 4
Number of data fields in collection or observation record	60 Notes: Varies a lot. Examples include Current name, name as received, What host or substrate, Locality, Who received from, Chain going back of who held it previously and those collection numbers, Kind of organism, Quarantine states in NZ, Diagnostic comments, Cultural peculiarities (how to grow it), Pathogenisity data – name of host its been tested on, Related cultures.
Proportion of fields per record that could easily be exported to GBIF	25% Notes: About 15 fields in ICMP are Darwin Core related and about 50 would be ABCD related.
Data Quality	Rank: 4 Notes: Anything with uncertainty is annotated.
Data Dictionaries	Dictionaries used: Names module for fungi and bacteria Notes: Names module for fungi and bacteria. If a person can't find a match there is a process for entry of new names into the names module. This includes recording the citation and authors. The Names module will be integrated in with the plant names system as a move to a more sophisticated platform in the future. Only broad localities are recorded so place names are not important.
Connectivity	Network Connectivity Rank: 0 Notes: Ask Jerry Data Connectivity Rank: 0 Notes: Ask Jerry, its in an Access DB currently

Legal/IP Constraints	Rank: 4 Notes: Some data is deposited while the work is in progress, this would be suppressed until the work (including sometimes identification) is complete. Some of the data is used in work doing bio prospecting for companies, so this would also be suppressed. Apart from that ICMP was set up as an international repository so its purpose is to be freely available. The price for cultures is a service charge. IP Policy: Some specific policy relating to Bio prospecting work
Organisational Willingness	Rank: 4 Notes: external funding is the big thing here
Other Notes:	

5.4.8 Landcare Research - New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC)

Title	New Zealand Arthropod Collection, New Zealand Nematode Collection and Specimen and Information Database (NZAC)
Abstract	NZAC has the most complete coverage of terrestrial invertebrates of all the collections held in New Zealand. In addition to its fundamental systematics value, the collection underpins quarantine and border control decisions e.g., verifying the presence or absence of species in New Zealand for ERMA, or for confirming identity of newly arrived potential pests for MAF. There are about 6.5 million specimens, of which about 1 million are pinned.
Contact Person	Trevor Crosby
Organisation	Landcare Research
Temporal Extent	1880 to present, majority of records 1920s (a lot of types were described then when Cawthron set up) and 1960 – 1990 (setting up of the systematics group and publication of fauna of new zealand series.
Spatial Extent	98% New Zealand, others types from overseas for comparison types. Hold specimens on behalf of Pacific Island countries, majority aren't digitised, from SPC pest surveys in 1970s. 3-5% digitised records holding institution is not NZAC, the physical specimen may not be located in NZAC, to say that it has been studied. These are unlikely to be recorded in other NZ biodatabases, and holding institution is identified in NZAC DB. Normally done as part of a revisioning project.
Volume	Total Number of Collections or Observations – 6,000,000 Number of Collections or Observations in Database – 60,000 Number of Taxa in Database - 6,000
Ability to export the mandatory fields required by GBIF	Rank – 4
Number of data fields in collection or observation record	20 Notes: Varies with the group for things like locality, map grids, altitude, for names may or may not have higher levels of taxonomic data. Are for some authority, date, page number described, links into the literature – usually relate to the types. Also image (illustrations and photos). Only one major set of fields.

Proportion of fields per record that could easily be exported to GBIF	80% Notes: Some of the fields might need splitting, e.g. latitude – degrees and minutes are one field rather than two.
Data Quality	Rank: 3 Notes: Between 3 and 4. Done lots of checking of quality of the data post entry, e.g. sorting to see if wrong ones come out, plotting things, asking why. No defined certainty statements but checked through by specialists. Digitising isn't just a simple exercise of transcribing data, also requires lots of understanding (e.g. based on someone's handwriting, knowing where the collector had been, e.g. Rau, Ranui example). So double entry would be useful, but expert review is more important. A field that says 'how much can you trust this information' – e.g. an expert identification, reputable person, can't guarantee anything. Plan to do this for delivery of some information for TFBIS. Category B rather than category A for example. People generally need the information that goes around it. Getting QA done can work really well by making the data available and people externally checking it and giving feedback.
Data Dictionaries	Dictionaries used: NZ Gazeteer Notes: For geographic names use NZ gazeteer. There is a names module, based on what was done for Species 2000. Have a separate field for recording verbatim/as it is, but primary interpretation is from a standard list. The set of names will be provided through NZBugs portal so other can use it (similar to plant names).
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: ?? Notes: Ask Jerry or Mark about frequency.
Legal/IP Constraints	Rank: 3 Notes: 4 or 3. Current research projects won't be released until publication, information may be entered but there is uncertainty about the names until it's checked and revised. Don't think there are biosecurity/economic issues – e.g. MAF records aren't delivered on the Internet. But wouldn't release work in progress that might have these economic/biosecurity issues. Flag private/public. Some would be considered rare/endangered – location information isn't down to the precision that you can do with current GIS, only to closest km. Some minor IP issues from some researchers but only a few years, then normally public. IP Policy: CRI Act, LCR policy, NSD policy document for all collections, nothing special for NZAC.
Organisational Willingness	Rank: 4 Notes: external funding is the big thing here

Other Notes:	Publishing on the web as distinct from scientific papers in journals. The basic 'bread and butter' documents which aren't ground breaking research but are still useful. To publish a lot of this in print (especially true for images) would be prohibitively expensive and wouldn't get the circulation. To publish for new species had to be in print or on CD. Now Zootaxa is done by Zhi-Qiang Zhang, online peer reviewed journal. http://www.mapress.com/zootaxa/ Started in China because of cost structures of publication there. Will get the status as with regular journals. Grown rapidly. Avenues could be done in a similar way for publishing databases. So if a journal was started for informatics, or metadata for informatics this would have the same status. Journal publishing is 'evidence that it's been done', for a person's career, number of publications is a career measure.

5.4.9 Landcare Research - National Vegetation Survey Databank (NVS)

Title	National Vegetation Survey Databank (NVS)
Abstract	The National Vegetation Survey Databank (NVS - 'Nivs') is a physical archive and computer databank containing records from approximately 45,000 vegetation survey plots--including data from over 12,000 permanent plots. NVS provides a unique record, spanning more than 50 years, of indigenous and exotic plants in New Zealand's terrestrial ecosystems, from Northland to Stewart Island and the Kermadec and Chatham islands. A broad range of habitats are covered, with special emphasis on indigenous forests and grasslands.
Contact Person	Rob Allen
Organisation	Landcare Research
Temporal Extent	1950 to present
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations - 0 Number of Collections or Observations in Database – 1,260,000 Number of Taxa in Database – 1,616
Ability to export the mandatory fields required by GBIF	Rank: 4 Notes: 90% of data from NVS is in an interim system. 10% is still in and old, old system. There is a plan to get all the data into final stable system. This will have a direct connection to GBIF and to Bioweb. All public domain data in the interim system has already been put into GBIF.
Number of data fields in collection or observation record	Per Plot: 21 attributes (including spatial co-ordinates) Per Species: 5 attributes Notes: Species co-occurrence data, measures of abundance, geographic data, site attribute data (10 or 12 fields)
Proportion of fields per record that could easily be exported to GBIF	60% Notes: All can be extracted, although GBIF is probably only interested in a smaller group. Most won't fit into GBIF, however the small subset that would could easily be exported. For plot data perhaps just name/number and location, for species data 3/5 fields.

Data Quality	Rank: 3 Notes: No vouchers, and the quality is variable. Quality depends on who did the observation, how difficult the taxa are, instances where the nomenclature are not up to date and are just as entered. Certainty is variable e.g. tagged trees have no defined certainty statements. Broad certainty statements could be made against specific taxa. For species names codes there is a look up table. There are various location range checks and altitude checks. There is no proofing or double entry. There are some logic checks and informal spot checks are done. There is no versioning or audit trail, although some paper records of changes on datasheets are kept.
Data Dictionaries	Dictionaries used: Landcare Research Plant Names Database, some internal data dictionaries Notes:
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 0 Notes:
Legal/IP Constraints	Rank: 3 Notes: Some would need permission in terms of IP or private land. At the moment there are only two options in our IP policy for NVS – 1. data that is freely available, 2. data for which you have to ask permission to get it entirely or any piece of it (i.e. to get into those datasets at all). We would like to introduce a new level for certain kinds of queries (i.e. for a limited subset that was not related to the kind of analysis the data collector wanted to do). For example Susan Wiser has collected data from banks peninsula for research on outcrops. Many species occurrence, environmental, and ecological factors were measured to ask scientific questions about how outcrops work ecologically. Someone contacted her who was doing research across 6 major tree species nationally. They wanted that data which was fine, as it was effectively an 'orthogonal slice of the data' in relation to the IP value/research purpose it was collected for. It would be good to have the freedom to allow those types of uses without having to go back and check every time with the collectors of the data. IP Policy: Yes, this is formally stated on http://nvs.landcareresearch.co.nz/html/NVSprotocol.aspx. The protocol does not yet cover database to database access.
Organisational Willingness	Rank: 3 Notes:
Other Notes:	The observation and species/taxa numbers are from the top few ranked percentage of taxa occurrence data in recce records showing no particular species contributing over 3%. If we included tree diameter data then we would have a considerable number of additional records for a few species of trees which would dominate both the number of 'observations' and dominance of recorded taxa in the combined observaton set. If we included other kinds of data we have, might add say 10-15% records, and if we included data not currently in the public domain, (generally just a matter of getting permission) this would add another 30%.

5.4.10 Landcare Research - Mammal Observation Database

Title	Mammal Observation Database
Abstract	A database of point and known range observation data for Wallabies, Feral Goats, Chamois, Deer (6), Pigs, Thar in New Zealand
Contact Person	Wayne Fraser
Organisation	Landcare Research
Temporal Extent	1996 to present
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations - 266 Number of Collections or Observations in Database – 266 NB these figures represent the total point observations over a six year period. It is known that some of these species have during this time been eradicated from the observed location. Number of Taxa in Database - 16
Ability to export the mandatory fields required by GBIF	Rank: 4 Notes: Location would be difficult as it is range data
Number of data fields in collection or observation record	20 Notes: 15 – 20 fields on evidence details (e.g. visual observation), origin of animals, disease status, reliability rating, supplementary information
Proportion of fields per record that could easily be exported to GBIF	60% Notes:
Data Quality	Rank: 4 Notes: Uses a reliability rating for records. All existing range data is checked and rechecked by DOC and Regional Councils
Data Dictionaries	Dictionaries used: None Notes: Manual checking against topo maps
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 3 Notes: Data is updated quarterly for DOC, no need to do more often than that as there is not sufficient change
Legal/IP Constraints	Rank: 4 Notes: Some location data may not be able to be made publicly available due to the risk of hunters trespassing on private land. IP Policy:
Organisational Willingness	Rank: 3 Notes: The data will quite likely sit under DOC's NHMS and could be connected to GBIF through that.

Other Notes:	System is currently an Access database & ArcView GIS. A TFBIS funded project is being initiated in April 2005 to make these data web accessible. It is intended that DOC and selected Regional Council staff will be able to enter data into the system.

5.4.11 Landcare Research - 5 Minute Bird Counts Database

Title	5 Minute Bird Counts Database
Abstract	A database currently underdevelopment of 5 minute bird count data across New Zealand, primarily done before and after 1080 poison possum controls
Contact Person	Eric Spurr
Organisation	Landcare Research
Temporal Extent	1977 to present
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations - 30,000 five minute bird counts, hundreds of thousands of individual bird observations Number of Collections or Observations in Database – 30,000 five minute bird counts, hundreds of thousands of individual bird observations Number of Taxa in Database - 35
Ability to export the mandatory fields required by GBIF	Rank: 1 Notes: Currently not accessible but plans underway to make it so
Number of data fields in collection or observation record	15 Notes: Location (eg Motautau forest), sometimes grind reference circuit of lines or can be derived, Date, Observer (generally just initials but supporting), Station no., Time, Weather (Sun, Wind, Precipitation), Bird species (number counted in 5 mins, and whether heard or seen), Some additional notes (e.g. general vegetation information, altitude of station)
Proportion of fields per record that could easily be exported to GBIF	70% Notes:
Data Quality	Rank: 3 Notes: Training mechanisms are in place, some quality control methods through individual trials balancing across a range of people to avoid bias.
Data Dictionaries	Dictionaries used: None Notes: May be included in new database
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 1 Notes: Data is currently in Paradox and Excel. DOC will be creating a database of 5 minute bird counts (primarily forest bird counts), so this data may end up in there, and be connected to GBIF that way.

Legal/IP Constraints	Rank: 2 Notes: Some counts done for private clients (AHB, Fletcher Forests, Electricity companies), so would have to clear with them. A lot is not published so would want to have caveats. Very few would have no access, quite a few would require permission. IP Policy:
Organisational Willingness	Rank: 2 Notes: Landcare Research would probably not be prepared to fund it, would have to be totally funded or majority funded to do this.
Other Notes:	

5.4.12 NIWA – FBIS

FBIS encompasses a number of distinct datasets. Responses for survey questions in relation to data export, quality and connectivity cover all these datasets. Metadata for each individual dataset is provided following this table.

Ability to export the mandatory fields required by GBIF	Rank: 4
Number of data fields in collection or observation record	140 Notes: Up to 140 attributes per observation (i.e. this the total possible number in FBIS)
Proportion of fields per record that could easily be exported to GBIF	30%
Data Quality	Rank: 3
Data Dictionaries	Dictionaries used: None
Connectivity	Network Connectivity Rank: 5 Data Connectivity Rank: 4
Legal/IP Constraints	Rank: 3 Notes: We would need to be sure we were not creating a commercial disadvantage
Organisational Willingness	Rank: 3 Notes: NIWA would be somewhere between 3-4

5.4.12.1 NIWA - FBIS Aquatic Weed Data

Title	FBIS Aquatic Weed Data
Abstract	Comprises over 2,000 invasive plant survey results held by the Department of Conservation, New Zealand herbaria, and regional councils. The Pest Status and Aquatic Weed Risk Assessment Ranking of invasive plant species are also recorded.
Temporal Extent	January 1887 - present
Spatial Extent	New Zealand waterways

Volume	Total Number of Collections or Observations – 2,130 Number of Collections or Observations in Database – 2,130 Number of Taxa in Database – 44

5.4.12.2 NIWA - FBIS Benthic 2000 Data

Title	FBIS Benthic 2000 Data
Abstract	Macroinvertebrate abundance data from the littoral zone of a number of large New Zealand lakes
Temporal Extent	28/02/1967 - 02/02/2001
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations – 8,400 Number of Collections or Observations in Database – 8,400 Number of Taxa in Database – 105 (across all FBIS macroinvertebrate data)

5.4.12.3 NIWA - FBIS Lake Data Macrophytes

Title	FBIS Lake Data Macrophytes
Abstract	Comprises plant survey results for 206 lakes, using survey method of Clayton (1983), and other supporting information
Temporal Extent	21/02/1981 - 13/03/2003
Spatial Extent	New Zealand lakes
Volume	Total Number of Collections or Observations – 17,500 Number of Collections or Observations in Database – 17,500 Number of Taxa in Database – 148

5.4.12.4 NIWA - FBIS Lake Water Quality Data (Algae)

NB for the purposes of this document a number of FBIS Algae datasets from different site records have been combined into one metadata summary.

Title	FBIS Lake Water Quality Data (Algae)
Abstract	These data incorporate the results of water quality monitoring of lakes in New Zealand by NIWA and several Regional Councils for the purpose of state of the environment reporting
Temporal Extent	1998 to 2001
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations – 970 sites Number of Collections or Observations in Database – Periphyton data from 970 sites Number of Taxa in Database – 551 (across all FBIS Algae Data)

5.4.12.5 NIWA - FBIS New Zealand Freshwater Fish Data

Title	FBIS New Zealand Freshwater Fish Data
Abstract	Data include the site location, the species present, their abundance and size, as well as information such as the fishing method used and a physical description of the site. The latter includes an assessment of the habitat type, substrate type, available fish cover, catchment vegetation, riparian vegetation, water widths and depths, and some basic water quality measures.
Temporal Extent	1920 - the present
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations – 57,100 Number of Collections or Observations in Database – 57,100 Number of Taxa in Database – 59

5.4.12.6 NIWA - FBIS Stream Invertebrates Data

Title	FBIS Stream Invertebrates Data
Abstract	Macroinvertebrate abundance data from a number of New Zealand rivers & streams
Temporal Extent	1967 - 02/02/2001
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations – 16,700 Number of Collections or Observations in Database – 6,700 Number of Taxa in Database – 105

5.4.12.7 NIWA - FBIS Algae Data

NB for the purposes of this document a number of FBIS Algae datasets from different geographic areas have been combined into one metadata summary.

Title	FBIS Algae Data
Abstract	These data are based on samples of algae collected from rivers, small lakes, tarns and mire pools in New Zealand
Temporal Extent	1997 to 2003
Spatial Extent	New Zealand
Volume	Total Number of Collections or Observations – 19,710 Number of Collections or Observations in Database – 19,710 Number of Taxa in Database – 551 (across all FBIS Algae Data)

5.4.13 Department of Conservation - Bioweb Herpetofauna

Title	Bioweb Herpetofauna
Abstract	National records of herpetofauna species - distribution and abundance. Records of observations, includes measurements, population estimates, population home range, some DNA samples. Collected to understand the state of quality and quantity of Herpetofauna.
Contact Person	Andrew Townsend
Organisation	Department of Conservation
Temporal Extent	Historic to present
Spatial Extent	NZ and subantarctic
Volume	Total Number of Collections or Observations – 30,000 Number of Collections or Observations in Database – 15,000 Number of Taxa in Database - 60
Ability to export the mandatory fields required by GBIF	Rank: 3 Notes:
Number of data fields in collection or observation record	24 Notes:
Proportion of fields per record that could easily be exported to GBIF	90% Notes:
Data Quality	Rank: 5
Data Dictionaries	Dictionaries used: Uses Journal of Herptofauna, and specialist panel convened by DOC, species nomenclature as a module in Bioweb, topographic data for place names
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 4
Legal/IP Constraints	Rank: 4 IP Policy: No
Organisational Willingness	Rank: 3 Notes: Primarily a funding issue, there is some policy intent to be international players.

5.4.14 Department of Conservation - Bioweb Threatened plants

Title	Bioweb Threatened plants
Abstract	National records of threatened plant species - distribution and abundance, identification including pictures and alternative names.
Contact Person	Andrew Townsend
Organisation	Department of Conservation
Temporal Extent	historic to present
Spatial Extent	NZ economic zone.
Volume	Total Number of Collections or Observations - 3,700 Number of Collections or Observations in Database – 3,700 Number of Taxa in Database – 2,500
Ability to export the mandatory fields required by GBIF	Rank: 3
Number of data fields in collection or observation record	32 Notes:
Proportion of fields per record that could easily be exported to GBIF	90% Notes:
Data Quality	Rank: 5
Data Dictionaries	Dictionaries used: Uses Landcare standards for names, and a specialist panel convened by DOC for threat status.
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 3
Legal/IP Constraints	Rank: 3 IP Policy: No
Organisational Willingness	Rank: 3 Notes: As above

5.4.15 Department of Conservation - Bioweb Weeds

Title	Bioweb Weeds
Abstract	National records of weed species - distribution and abundance, identification including pictures, alternative names, and control techniques.
Contact Person	Clayson Howell
Organisation	Department of Conservation
Temporal Extent	Historic to present
Spatial Extent	NZ economic zone.
Volume	Total Number of Collections or Observations - 85,000 Number of Collections or Observations in Database – 85,000 Number of Taxa in Database – 2,400

Ability to export the mandatory fields required by GBIF	Rank: 3
Number of data fields in collection or observation record	8 Notes: 7-8 for observations
Proportion of fields per record that could easily be exported to GBIF	90% Notes:
Data Quality	Rank: 5
Data Dictionaries	Dictionaries used: Uses published botanical references
Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 5
Legal/IP Constraints	Rank: 5 IP Policy: No
Organisational Willingness	Rank: 3 Notes: As above

5.4.16 Department of Conservation - Bioweb Bird banding

Title	Bioweb Bird banding
Abstract	The National banding scheme keeps records of all birds banded in New Zealand and all recoveries.
Contact Person	Graeme Taylor
Organisation	Department of Conservation
Temporal Extent	early 1950's to present
Spatial Extent	For banding NZ economic zone, recoveries from anywhere in the world.
Volume	Total Number of Collections or Observations – 1,370,000 Number of Collections or Observations in Database – 1,250,000 Number of Taxa in Database – 250
Ability to export the mandatory fields required by GBIF	Rank: 2
Number of data fields in collection or observation record	34 Notes:
Proportion of fields per record that could easily be exported to GBIF	90% Notes:
Data Quality	Rank: 5
Data Dictionaries	Dictionaries used: Uses NZ ornithological checklist

Connectivity	Network Connectivity Rank: 4 Data Connectivity Rank: 3
Legal/IP Constraints	Rank: 3 IP Policy: Yes
Organisational Willingness	Rank: 3 Notes: As above

5.4.17 Canterbury Museum Zoology Collection

Canterbury Museum has a large zoology collection. A proportion of this is currently databased in their collection management system. For the purpose of this survey a distinction has been made between their vertebrate and invertebrate collections, even though they use the same database. Responses for survey questions in relation to data export, quality and connectivity cover both these datasets. Metadata for each of the datasets is provided following this table.

Ability to export the mandatory fields required by GBIF	Rank: 4 Notes:
Number of data fields in collection or observation record	12 Notes:
Proportion of fields per record that could easily be exported to GBIF	66% Notes:
Data Quality	Rank: 3 Notes: We are currently at 3 for all data entered some areas of data are aiming to achieve standard 4 but are some way short of this milestone
Data Dictionaries	Dictionaries used: NZ Geographic Places database and various specific taxonomic databases (mostly in the ITIS framework) Notes:
Connectivity	Network Connectivity Rank: 1 Notes: Data Connectivity Rank: 3 Notes:
Legal/IP Constraints	Rank: 3 Notes: But the phrasing of this question doesn't quite suit the Museum environment - Our major IP issue is that the museum gets acknowledged and that it is known that a specimen exists in the museum) IP Policy: Yes but again the major issue is acknowledgement
Organisational Willingness	Rank: 2 Notes:

5.4.17.1 Canterbury Museum - Vertebrate Zoology Collection

Title	Vertebrate Zoology Collection
Abstract	Database of Cantebury Museum’s collections including birds, mammals, fish, and subfossil birds and mammals

Contact Person	Dr Paul Scofield
Organisation	Canterbury Museum
Temporal Extent	1860 to present
Spatial Extent	In descending order of importance NZ, Australia, Antactica, Europe, USA, South America Indonesia, New Guinea, Russia
Volume	Total Number of Collections or Observations – 48,100 Number of Collections or Observations in Database – 41,200 Number of Taxa in Database – 2,800 NB The vertebrate collection also includes marine fish, but these numbers have been excluded for the purpose of this survey.

5.4.17.2 Canterbury Museum - Invertebrate Zoology Collection

Title	Invertebrate Zoology Collection
Abstract	Database of Cantebury Museum's invertebrate collection
Contact Person	Dr Paul Scofield
Organisation	Canterbury Museum
Temporal Extent	1860 to present
Spatial Extent	
Volume	Total Number of Collections or Observations – 250,000 Number of Collections or Observations in Database – 15,000 Number of Taxa in Database – 3,000

5.4.18 Te Papa - Natural Environment Collection

Te Papa's Natural Environment Collection encompasses a number of distinct datasets. Responses for survey questions in relation to data export, quality and connectivity cover all these datasets and are listed in a single table below. Metadata for each individual dataset is provided following this table.

Title	Natural Environment Collection
Abstract	Te Papa's natural history collections database, including plants, birds, molluscs, land mammals, reptiles and amphibians, fish, and arthropods. Only a percentage of specimens are databased, with percentages varying per collection. Collections are databased in Te Papa's collection management system, which is currently undergoing significant upgrade.
Contact Person	Patrick Brownsy
Organisation	Te Papa
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations - 922,400 (600,000 of which are arthropods, 250,000 are plants) plus 110,000 lots in the Molluscs collection. Number of Collections or Observations in Database - 191,300 plus 30,000 lots in the Molluscs collection Number of Taxa in Database - 0

Ability to export the mandatory fields required by GBIF	Rank: 4 Notes: A new Collection Management System is currently being implemented. This will be in by June 2005 and stable three to six months following that.
Number of data fields in collection or observation record	100
Proportion of fields per record that could easily be exported to GBIF	90% Notes: 85 out of 94 elements are mapped to Darwin Core fields. Have considered ABCD but it is to complicated for our needs.
Data Quality	Rank: 3 Notes: Will implement a "confirmed by" field. Could assign certainty statements but it would be costly, some data sets quality control is good, some not. Changes to major fields such as 'Identification' are logged. There are no datasets that stand out as being much lower in quality than other data sets, but it would be fair to say that the mollusc data is of exceptionally high quality and the plant data of very high quality.
Data Dictionaries	Dictionaries used: Names (internal), LINZ, Getty Institute Notes: Internal hierarchies of names as data dictionary. Place names – will use LINZ and Getty Institute thesaurus for international names. Use NZMS260 mapping sheets. Databases of collectors used as a data dictionary. Could use Landcare Research's list of plant names but only in batch delivered mode.
Connectivity	Network Connectivity Rank: 5 Notes: The planned Collections Online service will assume 24x7 access Data Connectivity Rank: 5 Notes: The planned Collections Online service will provide near real time access to collections information so this would effectively be the same for any connection to GBIF
Legal/IP Constraints	Rank: 4 Notes: Some restrictions in terms of safety for rare and threatened species. Some restrictions for scientists IP for research currently underway. IP Policy: A principal of making it available but nothing formal. Some data sharing agreements with DOC.
Organisational Willingness	Rank: 3 Notes: - Willingness has not been tested yet. There are implications in that more collection loan requests would be likely and there are associated cost implications. We only have a small IT team. We would have to have initial internal discussions, however most people, including executive management, would in principle want to provide data to GBIF.

5.4.18.1 Te Papa - Birds Collection

Title	Birds Collection
Abstract	Te Papa houses the world's greatest collection of New Zealand birds, numbering over 70,000 specimens. Highlights from this collection include many irreplaceable specimens of extinct New Zealand birds and one of the world's largest collections of oceanic birds. The main focus of the collection is on New Zealand species. New Zealand's land and freshwater bird fauna is comprised of just over 200 species, many of which are recent introductions or vagrants - birds that have strayed from their usual range.
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations - 64000 Number of Collections or Observations in Database - 64000 Number of Taxa in Database - 0

5.4.18.2 Te Papa - Fishes Collection

Title	Fishes Collection
Abstract	The national Fishes Collection held by Te Papa is by far the largest in the world from the New Zealand region of the Southwest Pacific Ocean. First started in 1869, it now comprises over 42,000 catalogued lots, each lot containing one or more specimens. These include over 1700 type specimens, the original specimens on which published descriptions of new species are based. Around 5% of catalogued specimens are freshwater.
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations - 2100 Number of Collections or Observations in Database - 2000 Number of Taxa in Database - 0

5.4.18.3 Te Papa - Land Mammals Collection

Title	Land Mammals Collection
Abstract	A collection of terrestrial mammals, including extensive and important collections of introduced rats, hares, rabbits, wallabies, mustelids (weasels, stoats, etc), and possums, as well as rare native bats, and the skeleton of the famous racehorse Phar Lap.
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations - 1300 Number of Collections or Observations in Database - 1300 Number of Taxa in Database - 0

5.4.18.4 Te Papa - Molluscs Collection

Title	Molluscs Collection
Abstract	The Mollusca Collection at Te Papa is the largest and most comprehensive in the country and comprises about 335,000 sample lots and several million specimens. Its scope is worldwide, with its greatest strength in specimens from the New Zealand region, including the Kermadec Islands, Norfolk Island, the subantarctic islands, and the Antarctic. Around 33% of molluscs in the collection are terrestrial.
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations – 110,000 lots (lots contain from a few to over 100 specimens) Number of Collections or Observations in Database –30,000 lots Number of Taxa in Database - 0

5.4.18.5 Te Papa - Plant Collection

Title	Plant Collection
Abstract	Te Papa maintains a Plant Collection, or herbarium, of almost 250,000 dried specimens. It covers both native and introduced flowering plants and gymnosperms, ferns, mosses, liverworts, lichens, and marine algae, from all parts of New Zealand. As well as dried specimens there are plant fossils, a range of timber samples, and material preserved in alcohol.
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations – 250,000 Number of Collections or Observations in Database – 110,000 Number of Taxa in Database - 0

5.4.18.6 Te Papa - Reptiles & Amphibians Collection

Title	Reptiles & Amphibians Collection
Abstract	The Reptile and Amphibian Collection contains approximately 6000 lots of reptiles and frogs from the New Zealand mainland and offshore islands, as well as some fossil marine-reptile material and foreign specimens. The collection is a permanent storage facility for specimens, many of which are rare and endangered, and is used for the comparison and identification of existing and new species.
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations – 5,000 Number of Collections or Observations in Database – 5,000 Number of Taxa in Database - 0

5.4.18.7 Te Papa - Arthropod Collection

Title	Arthropod Collection

Abstract	The collection is particularly strong in beetles, butterflies, moths, lice, fleas, stick insects, springtails, cicadas, wētā, spiders, harvestmen, and water bears. It includes about 1100 primary types - the original specimens on which published descriptions of a species are based - within an estimated 600,000 specimen lots. Although Te Papa's collection focuses on the New Zealand subregion of the world, there are also sizeable quantities of material from Australia, the Pacific Islands, and other places.
Temporal Extent	1769 till present day
Spatial Extent	New Zealand, some Pacific collections
Volume	Total Number of Collections or Observations – 600,000 Number of Collections or Observations in Database – 9,000 Number of Taxa in Database - 0

5.4.19 Otago Museum - Natural Environment Collection

Title	Natural Environment Collection
Abstract	Database of Otago Museum's collections including birds, vertebrates, native plants, terrestrial invertebrates, insects and spiders.
Contact Person	Brian Patrick
Organisation	Otago Museum
Temporal Extent	1880 to present
Spatial Extent	New Zealand including sub Antarctic, pacific, strong on parts of Asia and Australia, patchy in Europe, North America, poor on Africa. It is part of our collection policy to have collections from other areas as it is good for research (for comparison), and for education purposes.
Volume	Total Number of Collections or Observations – 2,100,000 () Number of Collections or Observations in Database – 210,000 (about 10% of the collection in total is digitised, however the majority is arthropods which skews the figures, more like 80% of the non arthropod collections are in the database) Number of Taxa in Database – 15,000 (again mostly invertebrates).
Ability to export the mandatory fields required by GBIF	Rank: 5 Notes:
Number of data fields in collection or observation record	30 Notes: 8 to 12 typically, for some up to 30. We are also putting in images and publications including scanned hand written notes.
Proportion of fields per record that could easily be exported to GBIF	80% Notes: The majority should be easily exportable, we made sure of this before we purchased our collection management system (Vernon).

Data Quality	Rank: 4 Notes: Will be a 5 by the end of the year. Audit NZ and the Ministry of Culture and Heritage are coming to do a full audit of our collections and databases.
Data Dictionaries	Dictionaries used: CRI floras and faunas of NZ, and gazetteers for place names Notes: All are internal copies based on CRI lists, journals, publications and web sites. We do systematics research and do change names and update them. We try to make sure these name changes are fed back to other organisations holding lists of taxonomic names.
Connectivity	Network Connectivity Rank: 5 Notes: Applicable as long as we could check all the security issues/implications. Data Connectivity Rank: 5 Notes: Our long term plan is to provide public access to all data in our system
Legal/IP Constraints	Rank: 5 Notes: There are minimal constraints anyway, and once something is put into the database the issues have been dealt with already.
Organisational Willingness	Rank: 4 Notes: On the high end of willingness, but external funding would help. We are likely to do it eventually anyway, but it will take longer without external funding.
Other Notes:	

5.4.20 Lincoln University - Centre of Research Excellence Database

Title	Centre of Research Excellence Database
Abstract	A database of terrestrial and aquatic insects collected for DNA analysis. The purpose of the research was to confidently attach DNA sequence data to individual specimens to facilitate DNA identification methods to meet New Zealand's biosecurity requirements. Exotic and endemic specimens were obtained from within New Zealand and around the world for this research.
Contact Person	Karen Armstrong
Organisation	Lincoln University
Temporal Extent	1920's to present
Spatial Extent	Global
Volume	Total Number of Collections or Observations – 3,000 Number of Collections or Observations in Database – 3,000 Number of Taxa in Database – 500
Ability to export the mandatory fields required by GBIF	Rank: 4 Notes: Used Darwin Core

Number of data fields in collection or observation record	18 Notes: Including sequencing data
Proportion of fields per record that could easily be exported to GBIF	60% Notes:
Data Quality	Rank: 2 Notes:
Data Dictionaries	Dictionaries used: None yet Notes: Will be hooked into a Landcare Research database for invertebrate names. No place names are recorded.
Connectivity	Network Connectivity Rank: 1 Notes: Will initially be on a separate server with a standalone PC – developed in conjunction with Sue Warner, invertebrate ecologist. Applied for TFBIS funding (with Jerry Cooper) but held up as someone in Auckland had a similar proposal but with Gen Bank Data Connectivity Rank: 3 Notes: Data will keep being added to. End user will be MAF (border and preborder) but they won't be the only user.
Legal/IP Constraints	Rank: 4 Notes: Could do all but aspects like gene sequence data for publishing reasons IP Policy: None yet
Organisational Willingness	Rank: 2

5.4.21 Massey University - Dame Ella Campbell Herbarium (MPN) Database

Title	Dame Ella Campbell Herbarium (MPN) Database
Abstract	Collection of mainly terrestrial higher plants (15,000 specimens), collected by past staff and students, as well as some donated collections (e.g. McEwen Coprosma collection). Extensive collection (15,000) of bryophytes, particularly liverworts (Hodgson and Campbell collections). Some misc pickles etc.
Contact Person	Dr GL Rapson, Ecology Group, Institute of Natural Resources
Organisation	Massey University
Temporal Extent	1874 to present
Spatial Extent	Focus on the lower North Island, New Zealand, but with specimens from around New Zealand, and a small collection of overseas' specimens.
Volume	Total Number of Collections or Observations – 30,000 Number of Collections or Observations in Database – 8,000 Number of Taxa in Database - 0

Ability to export the mandatory fields required by GBIF	Rank: 4 Notes: We don't know how to do this, but we think an expert would
Number of data fields in collection or observation record	19
Proportion of fields per record that could easily be exported to GBIF	80% Notes: 80% of non-mandatory fields probably exportable.
Data Quality	Rank: 3 Notes: Mixture – standard data entry quality and integrity control, reasonable methods to ensure data quality; ad hoc collection monitoring.
Data Dictionaries	Dictionaries used: Landcare Plant Names database Notes: Use Landcare Plant Names database, and misc. others for checking taxonomy
Connectivity	Network Connectivity Rank: 1 Notes: We could put a database on the web, but here are no plans to do so. GBIF query structure could not be set up without extensive negotiation with management of IT Data Connectivity Rank: 1 Notes: 1 - 2, Low volumes of new data, and low priority to update to GBIF
Legal/IP Constraints	Rank: 5 Notes: 5 is probably the answer, but I suspect that IP issues have never been addressed IP Policy: No policy to date, but the issue will arise
Organisational Willingness	Rank: 2 Notes: Low priority compared with databasing our existing collection
Other Notes:	We believe that the electronic database is there to support the physical collection, and is at no time a substitute for looking at the actual specimens.