**Dataversity**

**Biodata Management Framework**

Dan Randow, Working Draft, 30 November 2012. [CC BY 3.0](http://creativecommons.org/licenses/by/3.0) TFBIS project 263: Biodata Management Framework: Phase Two.

**Introduction**

**Assessing and Planning Improvements to Biodata Systems**

Efforts to manage biosecurity and protect biodiversity depend on data about species and ecosystems. Data is required for policy, planning and management within organisations. And it is required for coordinating activities between organisations.

Biodiversity and biosecurity data is complex and messy. There are many different aspects of managing it effectively to consider. When planning improvements in a specific area, it is hard to know what to focus on. When assessing and planning system-wide improvements, it is hard to know where to start. It is hard to know where to invest in improving systems to get the maximum benefit in terms of better data to support biodiversity and biosecurity management.

**A Framework to Facilitate Investment in Biodata Systems**

This Framework can be used to increase the return on biodiversity and biosecurity information system investments in all organisations. It can be used to assess the state of biodata systems, to set goals for improving systems, and to plan and measure improvements in specific areas.

Adoption of the Framework is intended to driven by improving the return on investment in biodata systems. By promoting and facilitating the adoption of standards and common approaches, use of the Framework will also improve the potential for data-sharing between organisations.

**Intended Uses of the Framework**

Biodata Managers will use the framework to:

- **assess** biodata systems,
- **prioritise** improvements to biodata systems,
- **plan** improvements to biodata systems,
- **evaluate** biodata systems prior to acquisition, and
- **measure** the success of investment in biodata system improvements.

Other potential uses include the following.

- Biodiversity Managers could use the Framework to assess the usefulness of datasets held internally or externally.
- Biodata System Providers could use the Framework to verify the maturity of their system.
- A catalogue of biodata systems could rate systems using the Framework, to make it easy for an organisation to find a system to meet its needs.
- In future, the Framework could be developed into a validated tool that can be used as part of the contractual basis for biodata system development and science delivery projects.

---

http://dataversity.org.nz  
dan@onlinegroups.net
**Scope of the Framework**

This Framework is for assessing systems to manage biosecurity and biodiversity data. This includes data about the following.

- **Species** – indigenous and pest species.
- **Ecosystems** – habitats for indigenous species.
- **Management** – activities carried out in the context of policies, programmes, projects, priorities and designations assigned to sites.

The scope of the Framework is independent of the following.

- Taxa and ecosystem types.
- Purpose for the data, including planning, management, operations, education and community engagement.
- Method used to collect the data, including casual observations, community monitoring, rigorous surveys, legacy datasets and remote data.
- Organisation managing the data, including government, NGOs and the private sector.
- Tool used to manage the data, including paper.

**Structure of the Framework**

The Framework divides up the task of biodata management into a set of Areas, each reflecting one stage in the *data life-cycle*.

The Framework defines five levels of *maturity* of data-management, determined by a set of *factors*. Those factors are summarised in a table showing how they are aggregated into maturity levels.

For each stage in the data life-cycle, the Framework defines *criteria* for five maturity levels. These are summarised in a System Maturity Matrix.

A System Assessment Form enables the Maturity Matrix to be used to assess the current maturity and set goals to improve a system.

The Framework provides a detailed guide to assessing and improving data-management at each of the four data-lifecycle stages: Capture, Ingest, Analyse and Share.

**Criteria for Biodata Management Areas**

These Areas are the second generation of Areas to be developed.

The criteria for well-defined areas, learned from the first generation are as follows.

- **Specific** – specific to biodata, or generic but relevant. Each handled appropriately.
- **Exclusive** – Clear boundaries. No overlap. Clear relationships.
- “Axis” – Should all either relate to a specific stage, or to all stages in the data life-cycle.
- **Size, Importance** – Each chunk should be about as large and important as the other chunks.
Data Life-cycle
There are four main stages in the biodata life-cycle:

1. **Capture** – Record data in the field.
2. **Ingest** – Introduce local or remote data into a data-management system and process it to meet system criteria.
3. **Analyse** – Produce summaries of data to answer specific questions.
4. **Share** – Make raw or summarised data available to people and other systems.

Maturity
Maturity refers to the fitness for purpose of data management systems. It maturity depends on various considerations and how well they are addressed. Maturity can be assessed consistently across all stages in the data life-cycle.

Maturity Criteria
The Maturity Matrix will contain a maturity criterion for each maturity level at each stage of the data life-cycle.

Goals of Maturity
As maturity increases, both of the following increase:

- **Effectiveness** – The production of useful data.
- **Efficiency** – The reduction of the costs of producing useful data.

Maturity is about Practices and Tools
Biodata management systems are made up of tools and practices.

- **Practices** – Processes that are followed to manage data.
- **Tools** – Technology that is used to manage data.

Maturity of systems requires maturity of both practices and tools.

Practices are used to achieve effectiveness, while tools are used to achieve efficiency.

At lower levels of maturity, practices are more important than tools. There is no point in using good tools with poor practices. Good practices can provide useful data, even if basic tools are used.

Maturity Levels
The framework will define five levels of maturity of biodata-management systems.

The maturity levels should permit consistent assessment of systems (tools and practices) across all stages of the data life-cycle. Each level should have a set of criteria, to facilitate assessment and planning to get from one level to the next.

In order to support prioritisation of investment in systems to support all stages of the data life-cycle, the criteria for maturity levels should be able to be applied consistently across all stages of the data life-cycle.
Structure of the Framework

At each level, the criteria for maturity in tools and practices must both be met. Maturity levels for practices have been well-researched. Established models for the maturity of tools are not as easy to find.

Factors that Determine System Maturity

The following factors determine the maturity of biodata systems.

• **Processes** – Data is consistently managed by well-maintained processes.
• **Business** – System is clearly licensed, reasonably priced, well-supported and well-maintained by a reliable and experienced vendor.
• **Platform** – Technical platform is robust, scalable, interoperable.
• **Reliability** – Ensure that available data is reliable enough to be correctly interpreted by its audience.
• **Permissions** – Ensure that data is only available to authorised viewers.
• **Standards** – Ensure that data is shared using standards that enable interoperability. These include metadata standards as well as taxonomic, geospatial and other data standards.

Processes

The Capability Maturity Model (CMM), developed originally by Carnegie Mellon University, has been adapted by April Reeve¹ for use with DAMA² and DMBOK³ for a Data Management Maturity framework.

**Level 1: Immature (Initial)**

The best practice activities are not performed by the organization. The best practice tools are not available or not used.

**Level 2: Repeatable (Repeatable)**

Some parts of organization are using recommended tools and processes while other parts are not.

**Level 3: Managed (Defined)**

The organization has a documented standard for performing the assessed activity or activities consistently and using applicable tools effectively.

**Level 4: Monitored (Managed)**

The process in question is established, tracked and monitored. Recommended tools are in place and being used consistently across the organization.

² The Data Management Association International (DAMA)
³ Data Management Body of Knowledge (DMBOK)
Structure of the Framework

Level 5: Continuous Improvement (Optimizing)
The activity is continually reassessed, improved upon, tracked and built into process.

Business
Mature data management tools satisfy the following business criteria.

- **Licencing** – tools are available cost-effectively under a licence that meets business needs.
- **Support** – support is available from a reliable source.
- **Maintenance** – regular updates are available.
- **Hosting** – Hosting provides adequate performance, security, integrity and archiving.

Platform
Mature data management tools satisfy the following platform criteria.

- **Proven** – Uses proven technologies with a good life-expectancy that is aligned with in-house policies.
- **Modular** – Architecture is scalable, flexible, extensible, modular and open.
- **Cross-Platform** – Has desktop, web and mobile interfaces.
- **Usable** – Uses web-standards, and follows accessibility and usability guidelines.
- **Interoperable** – Has machine interfaces (ie web services and an API). Supports integration with in-house and external systems that carry out specialised functions that are beyond its scope. Integration should handle authentication and permissions.

Licensing
Almost all biodata is of potential value in developing national and international understanding of biodiversity state and trends. Most biodata can be shared with the public without constraints.

In some cases, however it is necessary to restrict access to biodata. Data may be collected under an agreement with a land-owner that restricts sharing. Some data relates to the occurrence of species that could be the target of rare species trafficking.

The following permissions information should be associated with all datasets.

- Copyright Owner – some data is owned by clients, or other third parties.
- Data Licence – the uses of the data that are permitted by the copyright owner.

Types of licence include the following.

- Creative Commons.
- The NZ Government Open Access Licensing framework (NZGOAL) provides a framework for licensing open government information based on Creative Commons.
- Data commons.
- A licence restricting use to specified parties.
- A licence that restricts the use of granular data but allows wider use of the data in an aggregated form.
Format
The format in which data is stored determines the ease with which the data can be used and exchanged. Data in paper notebooks is not as easily used and exchanged as data available by standards-based web services.

Level 1: Non-Digital
Paper notebooks, photographic prints.

Level 2: Unstructured
Images, unstructured PDFs.

Level 3: Structured (proprietary)
eg Excel.

Level 4: Structured (non-proprietary)
eg CSV. Standard archives.

Level 5: Granular
Individually reference-able entities.

Reliability
The reliability of data depends on the method used to collect it, the rigour with which that method was followed. The reliability of data as an indicator of current state depends on those things and the age of the data. Data that is highly volatile (eg bird counts) loses reliability faster than data of low volatility (eg GPS references).

The system should allow data of varying reliability to be managed. The reliability of the data must be recorded when data is imported. The system should maintain data reliability indices as they change over time. It should make the reliability of the data easily discoverable. Only data of higher reliability should be shown to users who are not qualified to interpret data of low reliability.

The data depreciation model developed by Bay of Plenty Regional Council functions as follows.
Each dataset has a current reliability value denoting its value to non-specialists for determining current state. The system uses the current reliability value to determine the sort order of datasets or to filter out data of low value.

When each dataset is ingested, it is assigned an initial reliability value. A casual observation of a bird in a wetland has a lower reliability than a survey using a rigorous methodology.

The dataset is also assigned a depreciation profile that determines the rate, and changes to the rate at which a dataset depreciates over time. Data does not depreciate in a straight line. It may depreciate slowly for the first ten years, and then more quickly for the ten years after that.

Data depreciation profiles are determined by technical experts for specific dataset types depending on the species group, purpose, methodology and conditions of the collection. Datasets must be matched to the correct depreciation profile when they are imported.
Structure of the Framework

Current values for all datasets are recalculated nightly.

Standards

In order to facilitate data-sharing with other NZ agencies and via international mechanisms, the system should expose all for sharing in a form that complies with applicable international standards.

Species data standards:

• Species occurrence data – Darwin Core⁴.
• Exported datasets – Darwin Core Archive⁵.

Ecosystem data standards:

• Survey data – Ecological Metadata Language (EML)⁶.

Geospatial data standards:

• Include the NZGO Spatial Data Infrastructure⁷, Australia New Zealand Land Information Council (ANZLIC)⁸, Open Geospatial Consortium (OGC)⁹.

Metadata Standards:

• Freshwater data – The The Freshwater Biodiversity Interoperability Framework¹⁰ being scoped by NIWA (TFBIS project 282) aims to provide nationally agreed way to easily query or transfer freshwater biodiversity survey information across organisations and systems.
• Metadata - monolithic (self-contained data) v federated (centralised search facility harvesting data from numerous sources)
• Metadata provides a reference back to "point of truth" - source of data
• Enable listings in Data Catalogues such as http://dc.niwa.co.nz/niwa_dc/srv/en/main.home
• Include the OGC Catalog Service¹¹.

---

⁴ http://rs.tdwg.org/dwc/ and http://rs.gbif.org/core/dwc_occurrence.xml
⁶ http://knb.ecoinformatics.org/software/eml/
⁹ http://www.opengeospatial.org/standards
¹⁰ http://dataversity.org.nz/groups/interoperability
¹¹ http://www.opengeospatial.org/standards/cat
How Maturity Factors relate to Maturity Levels

Summary of how maturity with respect to specific maturity factors aggregates to determine overall system maturity.

<table>
<thead>
<tr>
<th>Factor</th>
<th>One</th>
<th>Two</th>
<th>Three</th>
<th>Four</th>
<th>Five</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processes</td>
<td>Immature (Initial) – The best practice activities are not performed by the organization. The best practice tools are not available or not used.</td>
<td>Repeatable (Repeatable) – Some parts of organization are using recommended tools and processes while other parts are not.</td>
<td>Managed (Defined) – The organization has a documented standard for performing the assessed activity or activities consistently and using applicable tools effectively.</td>
<td>Monitored (Managed) – The process in question is established, tracked and monitored. Recommended tools are in place and being used consistently across the organization.</td>
<td>Continuous Improvement (Optimizing) – The activity is continually reassessed, improved upon, tracked and built in to process.</td>
</tr>
<tr>
<td>Business</td>
<td>Ad hoc management of tools.</td>
<td>System integrity and backups managed.</td>
<td>Explicit software licensing and support arrangements.</td>
<td>Maintenance provided by established vendor.</td>
<td>System management integrated with all-of-business IT processes.</td>
</tr>
<tr>
<td>Licensing</td>
<td>No permissions recorded or enforced.</td>
<td>Broad permissions known, ad hoc management.</td>
<td>Explicit licences associated with datasets.</td>
<td>Permissions vary with aggregation.</td>
<td>Permissions dynamically integrated with all interfaces.</td>
</tr>
<tr>
<td>Reliability</td>
<td>No reliability information recorded.</td>
<td>Ad hoc management of reliability information.</td>
<td>Static quality indices used.</td>
<td>Data depreciation.</td>
<td>Reliability dynamically integrated with all interfaces.</td>
</tr>
<tr>
<td>Standards</td>
<td>No data standard used.</td>
<td>Proprietary data standard used.</td>
<td>Externally referenced data standard used.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# System Maturity Matrix

Mockup of the matrix for getting a feel for life-cycle stages, maturity levels and maturity criteria.

<table>
<thead>
<tr>
<th>Maturity Level</th>
<th>Lifecycle Stage</th>
<th>Capture</th>
<th>Ingest</th>
<th>Analyse</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five</td>
<td></td>
<td>Record data in the field.</td>
<td>Introduce local or remote data into a data-management system and process it to meet system criteria.</td>
<td>Produce summaries of data to answer specific questions.</td>
<td>Make raw or summarised data available to people and other systems.</td>
</tr>
<tr>
<td>Four</td>
<td></td>
<td>Integrated digital capture.</td>
<td>Changes are controlled with an audit trail &amp; roll-back.</td>
<td>System can carry out spatial analysis.</td>
<td>Data catalogue is visible externally. License and metadata available; format for national exchange</td>
</tr>
<tr>
<td>Two</td>
<td></td>
<td>Data captured using well-managed manual systems.</td>
<td>Basic known process.</td>
<td>Analysis follows documented processes.</td>
<td>Manual request and supply. Static documents (eg PDF) available online.</td>
</tr>
<tr>
<td>One</td>
<td></td>
<td>Data captured in a variety of uncoordinated ways.</td>
<td>No way to know what data exists or where it is.</td>
<td>Analysis is ad hoc.</td>
<td>Data not easily exchanged</td>
</tr>
</tbody>
</table>
System Assessment Form

<table>
<thead>
<tr>
<th>Organisation:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>System:</td>
<td></td>
</tr>
<tr>
<td>Assessed by:</td>
<td></td>
</tr>
<tr>
<td>Date:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Current</th>
<th>Three Year Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>Level</td>
</tr>
<tr>
<td>Capture</td>
<td>2</td>
</tr>
</tbody>
</table>

|  |  |  |  |  |
|---|---|---|---|
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
Capture

Record data in the field.

Description

What is important:

- Know what to observe, what data to capture.
- Access information about the site that is useful for the survey process.
- Capture data about the survey, as well as observations data. (Data captured about the survey should be sufficient to determine the reliability and permissions of the data.)
- Use standard organism names and ecosystem assessment criteria and notation.
- Capture data at a granular level (rather than aggregating any data in the field).

Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>What Happens</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>• Digital field capture with validation.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• GPS integrated into capture device.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Direct live or batch synchronisation of data with primary data repository.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>• Generic application on hand-held device.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Synchronisation via generic file formats.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Hand-held GPS unit.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>• Paper forms defining all data to be captured.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Site information.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>• Paper without forms. Notebooks.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Casual observations by qualified experts or the public.</td>
<td></td>
</tr>
</tbody>
</table>
**Ingest**

Introduce local or remote data into a data-management system and process it to meet system criteria.

**Description**

What is important:

- Original data set is retained. May be notebooks, paper forms, spreadsheets, CSV files, databases in obsolete file formats, or live datasets on remote systems.
- Data is validated (unless it is known that the data was validated during capture).
- Standards-compliance is ensured. Species names, eg should be obtained from NZOR.
- Metadata such as who collected the data, when and why is recorded.
- The reliability and permissions of the dataset is recorded.
- Dataset is authorised for ingestion.
- Authorisation and all other modifications to a dataset are recorded, with intact copies of the previous dataset available to support rollback or auditing.
- Local data is discoverable elsewhere, and remote data is discoverable locally.

**Levels**

<table>
<thead>
<tr>
<th>Level</th>
<th>What Happens</th>
<th>Benefits</th>
</tr>
</thead>
</table>
| 5     | - Data is ingested into a single system.  
       | - Changes to remote datasets can be carried out dynamically. |  |
| 4     | - Changes are controlled and tracked, with previous versions accessible.  
       | - Relevant data (including legacy and external) is catalogueed.  
       | - Data catalogue is integrated with external catalogues. |  |
| 3     | - New datasets are routinely ingested using controlled process.  
       | - Legacy data and external datasets can be catalogueed.  
       | - Standards-based data catalogue maintained.  
       | - Reliability and permissions information are recorded. |  |
| 2     | - Data catalogue is maintained.  
       | - Data is validated. |  |
| 1     | - No ingestion process.  
       | - Unknown number of systems. |  |
Analyse

Produce summaries of data to answer specific questions.

Description

What is important:

- Discovery: Find relevant data that exists, either locally or remotely. (People in other parts of the organisation can see the presence of relevant biod or bios data.)
- Determine the provenance of data.
- Distinguish duplicates of a dataset.
- Carry out ad hoc comparisons of two or more datasets to determine relationships between them. (For example, relate data about species or ecosystems to contextual data such as status under one or more designations, priority, or work programme.)
- Build and save queries for routine analyses.
- Use data at varying levels of granularity.
- Compare data using spatial criteria.
- Data reported against internally developed indicators.
- Data statistically interpolated; automatic calculation and reporting; visualisations
- Legends, layers, maps, interactive models.
- Data automatically reported in terms of authoritative standard indicators that are easily digestible by both public and scientists.
- system automatically calculates ecological priority.
- system integrated with a decision-support system.

Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>What Happens</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>• Stored queries access dynamic datasets.</td>
<td>•</td>
</tr>
<tr>
<td>4</td>
<td>• System can carry out spatial analysis.</td>
<td>•</td>
</tr>
<tr>
<td>3</td>
<td>• System can aggregate data.</td>
<td>•</td>
</tr>
<tr>
<td>2</td>
<td>• Analysis follows documented processes.</td>
<td>•</td>
</tr>
<tr>
<td>1</td>
<td>• Analysis is ad hoc.</td>
<td>•</td>
</tr>
</tbody>
</table>
Share

Make raw or summarised data available to people and other systems.

Description

What is important:

- data externally discoverable.
- Reports are easy to read.
- Raw data is shared in open formats, for either batch download or dynamic access via an API.
- It is possible to determine the original data sets that were used to compile a report.
- Permissions are enforced correctly.
- Only data of high reliability is shown to people who are not qualified to interpret data of low reliability.
- Raw data is shared in standards compliant form.
- Users of other in-house systems can discover the existence of biodata associated with sites they are dealing with.

Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>What Happens</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>• Data available via web services. Format suitable for international exchange</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>• Data catalogue is visible externally.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Licence and metadata available; format for national exchange</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>• Organisation or region-wide systems interoperable. Raw data available online.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>• Manual request and supply. Static documents (eg PDF) available online.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>• Data not easily exchanged</td>
<td></td>
</tr>
</tbody>
</table>