Atlas of Living Australia

GUIDE TO DATA QUALITY

Author(s): Miles Nicholls

Version: 1.2

Date: 10/2/2011

File: ALA Quality Guide v1_2.doc

Revision history

Version	Date	Author(s)	Change description
1.2	10/2/2011	Miles Nicholls	Transferred to template
1.1	12/1/2011	Miles Nicholls, Lee Belbin	Integrated comments from L.Belbin
1.0	11/1/2011	Miles Nicholls	Initial release

Table of contents

1. Introduction

1.1 Summary

ALA Data Management will support the collection and sharing of data with documented quality parameters through clearly documenting a target data and metadata model based on supplier and user needs. Systems will be developed to implement the process of data collection, digitisation, validation, cleaning and access.

1.2 The current situation

The Atlas currently shares over 20 million species occurrence records from institutions and collectors including herbaria, museums and conservation agencies. This is a small percentage of the data currently available and of the data yet to be collected. To enable the data to be used in a cohesive manner to report on Australian biodiversity a set of quality principles need to be developed and integrated into the Atlas data management processes.

Data that has already been collected varies from undigitised to standardised and available. Data needs to be digitised, validated and shared. Legacy data (a term for the data that has already been collected) is an invaluable body of knowledge providing an irreplaceable baseline for biodiversity research.

The Atlas also needs to support the collection of new data through the development of data entry tools that facilitate data quality.

The Atlas can provide a substantial benefit to the research and management community through the transparent aggregation and validation of biodiversity information. The data currently available from state conservation agencies, natural history collections and Birds Australia is accessed by multiple organisations, multiple times, cleaned multiple times and used for analysis multiple times. The ALA can potentially remove the duplication of effort through centralisation of a transparent process. It is important to leverage the knowledge and experience of those who are currently validating data and to make the process efficient.

1.3 Priorities

Systems to facilitate the collection of new data are the immediate priority. Legacy data is a vital and irreplaceable resource, but it will still be there in 6-12 months. There is a limited window to establish standards and tools for data entry and several projects have approached the ALA for input including citizen science groups, AusPlots and the Great eastern ranges initiative.

2. Quality

2.1 Integrity

Data integrity considerations consist of whether the record is cohesive in terms of the field contents and whether the information makes sense or is usable in a real world context. This can be considered at any of the steps in the lifecycle of a record – original source, production of an export, import into another system, downstream processing.

A record with good integrity will have data in all appropriate fields and the data will conform to best current practice standards. Data values should be within specified bounds. Unless data meets basic integrity criteria it should not be loaded and referred back to the source.

2.2 Usability

How information will be used determines what constitutes a measure of quality in a particular context. To service the widest range of applications, users should be able to evaluate the fitness for use, or "usability", of data. It is the users, rather than the ALA who needs to determine which data will meet their quality threshold. There is also the potential to improve the quality of some data parameters from their context with other parameters. For example, the accuracy of a location may be improved using locality descriptions and known ranges for a species

The usability of data is based on metadata as well as the core measures. E.g. the accuracy of the geospatial coordinates as well as the coordinates themselves. It is important to ensure that the metadata is available and able to be used as a filter when selecting data for inclusion in an analysis.

Wherever possible metadata should be collected when systems and records are created to ensure it is as accurate and complete as possible. For older systems and records where metadata is known but not digitised it should be entered. The most difficult situation is where there is no documented metadata, in this case efforts need to be made to collect and digitise it. If no original metadata can be found then it may be possible to derive some from other data and details of the records. It is vital that all derived metadata is flagged as such and the methods used to produce it are easily accessible.

2.3 Other quality and usability considerations

Persistent identifiers

The use of persistent and preferably globally unique identifiers prevents duplication across systems and allows different versions of a record to reference their source. If a source system changes its database platform or owner it then the records are still able to be referenced.

Licensing and attribution management

The terms of use associated with a dataset or record are also usability (if not necessarily quality) considerations. The license associated with a data set needs to be available as a filter term.

3. Quality principles

For references and resources on data quality in the biodiversity domain see the References section.

3.1 Data needs metadata

- Data and datasets need sufficient metadata to allow a user to determine it's fitness for use
- Qualitative – descriptive, detail
- Quantitative - accuracy, precision

3.2 Prevention is better than a cure

- It is cheaper and more effective to ensure data is entered correctly the first time than to find and fix errors
- It is still necessary to validate and clean data but it will not have as large an impact on the records

3.3 Capture and store exact values at the highest precision possible

- Reduce accuracy for display if needed
- Categorise or group additionally rather than instead of recording the exact figure

3.4 Use systems and interface design to facilitate quality with minimal overhead

- Use reference lists in interfaces (also need to be able to enter new / unknown)
- Enter repeated elements once and reuse

3.5 Validate and Clean

- Validate and clean data
- Validate and clean metadata

3.6 Feedback

- Feedback potential gaps and errors to the source (but don't expect them to be fixed immediately)
- Feedback gaps and errors into data entry processes and systems to prevent errors

3.7 Transparency and Traceability

- Maintain verbatim values - retain original values so they can be reprocessed with different rules if needed
- Users need to see what's been done to a record to be able to have confidence in it and know whether or not to use it

3.8 Embed quality in the management process not just the technology

- Include quality checkpoints at collection organisations E.g. New collecting events are only approved or finalised when metadata is completed

4. Implementation

4.1 Occurrence record processing

The stages an occurrence record goes through are outlined below.

- Record – capture the information
- Digitise – enter the information into an electronic system
- Mobilise – make the information available by electronic means
- Validate – check for gaps and errors
- Clean – fill gaps and fix errors (in the context of associated data where possible?)
- Use – access and analyse
- Feedback – report back to the source

The stages may occur in a different order depending on the tools and processes used.

- Record and Digitise may occur in the same step
- Mobilise may take place at any time after the record is Digitised
- Data needs to be Validated before it is Cleaned
- Validation and Cleaning may occur before or after Mobilisation
- There may be several points where validation rules are applied – record, digitise, validate

Both new and legacy data go through the same process but legacy data has already been collected and in some cases digitised and mobilised. This can be an advantage but may also require a repeat of phases of the process depending the location of errors.

4.2 What makes up a good record?

For records that will be used in analysis rather than as a description, there is a value for each of a set of core measures (without which there is effectively no record) and extensions providing context to the core set.

To assess usability, each of the core values needs to have metadata.

- An indication of the accuracy and precision
- Information on verification or evidence for the value and accuracy - how can I check the value or have a confidence in it?

A good record allows fitness for use to be determined.

The presence/absence of values in any of these groups as well as the values themselves should be available as filters on data. It is important to be able to determine the difference between "value not displayed" (possibly for sensitivity reasons), "no value available" and "0" for example.

As the key analysis data of the ALA is the occurrence record, what makes a good record? The core values of an occurrence record are What, Where and When.

What

- Value: species
- Precision: to what taxonomic rank has the identification been made
- Accuracy: a term indicating how much confidence there is in the identification
- Verification:
    - basis of record (one of: observation, photo, specimen, sound, footprint etc)
    - identification method (one of: book, taxonomic key, expert identifier)

Where

- Value: coordinates, locality, location description
- Precision: how precise is the way of measuring location (e.g. GPS coordinates reported to five decimal places)
- Accuracy: margin of error in the location (within a 100m square from the point)
- Verification: GPS, Map, accuracy reduced for sensitivity reasons, other

When

- Value: date, time
- Precision: to what level was the time recorded – time, day, month, year, decade, etc
- Accuracy: over what time period could the event have occurred (1 hour, 1 day, etc)
- Verification: survey trip dates, diary/notebook entry, trap placement period

The fields to record this information are, for the most part, available in Darwin Core. Any that are not available (date accuracy for example) will be raised for consideration as an addition to the ALA data model and Darwin Core itself.

This grouping of fields into values, precision, accuracy and verification types may be used as the context for documentation explaining how to use Darwin Core fields. It can also be used in the interface design for data entry and mapping tools. E.g. When recording a sighting please indicate the certainty of the identification (select from options), the basis for the identification (select from options) and how the identification was determined (select from options).

Data entry systems developed by the ALA should encourage (but not necessarily make mandatory – sometimes the information is not available) comprehensive resource metadata for data sets. Data entry tools should encourage the entry of usability metadata at record level or through the setting of defaults that apply to particular sets of records such as surveys, time periods or users.

4.3 New data

To facilitate quality records, new data requires:

- Data entry tools and processes that facilitate complete and accurate quality data and metadata collection and management
- Documentation and training on the benefits of quality principles, tools and processes
- Mobilisation, validation and cleaning

The key to successfully establishing quality data collection will be to minimise overheads. System interface and workflow design will be vital to the success of the tools. The entry of

quality data at the point of collection will minimise validation and cleaning required. Reuse of repeated information through the creation of templates will help facilitate this.

E.g. A template for a particular collecting methodology that includes all the repeated metadata

- Methodology
  - Collection methodology
  - Basis of record
  - Collector name
- Taxonomy
  - Species pick lists based on area checklists or taxonomic focus
- Identification
  - Identified by
  - References
- Location
  - Coordinate accuracy and precision
  - How location was measured (e.g. GPS make and model, map name and scale)

The information above may be systematically applied to records and updated if there is a difference with a specific record. The entire template can also be reused for future events.

4.4 Legacy data

To facilitate quality records, existing data requires:

- Digitisation (if not yet digitised)
  - Entry of quality data and metadata when digitising records
- Validation
  - Identification of error types and recognition methods
  - Review of existing data and metadata against quality model and measures
- Cleaning
  - Contact data sources to complete records
  - derivation of data and metadata unavailable from source
- Mobilisation either before or after validation and cleaning

Validation of legacy data aims to bring previously collected data up to current quality standards. This will be an ongoing task that ALA systems need to facilitate whilst making the data available for use.

Validation (quality review of the data) tools focus on identification of gaps in quality so they may be able to be addressed in the future. The second part of the process is filling or repairing gaps and errors. Missing data should first be completed by identification and update from the original source data. If the data is not available from the source it may able to be updated by an expert or through the application of algorithms and standard values. If no other values are available, fields should be flagged to indicate that efforts have been made but no value was found. Validation and update processes may be carried out at either the source or the ALA. If validation occurs on ALA infrastructure then the data is available and may be suitable for some analyses even if it is not complete.

The completion of records with source data or update by an expert will be a resource and expertise intensive task and will continue long beyond the current timeframe of ALA funding. Research and collecting institutions are consistently resource poor and the review and update of data is often a low priority. As a result the validation and record update systems developed by the Atlas should focus on identification and flagging of gaps and potential errors but not rely on the issues being immediately addressed at the source. The traditional goal of a closed feedback loop may not prove to be practical. ALA systems should aim to display multiple versions of records so that corrections are available even if they have not yet been updated at the source.

4.5 Data set and record metadata

Usability metadata exists at multiple levels including the collection, dataset and individual record level. There may also be grouping of records with data sets e.g. individual surveys with a particular methodology. The majority of metadata for the data set also applies to every record within that data set however it may be a requirement of a particular analysis that data is selected from data sets with particular characteristics rather than based on record characteristics.

Data set level metadata is particularly important for distinguishing survey or longitudinal data sets. The individual records are occurrences but they may be repeated observations of the same individual(s) over time. Knowledge of the data set as a whole is needed to make an assessment of its usability. These data sets can be extremely valuable for monitoring but incorrect use of these records can lead to misinterpretation. Presence / absence data sets, although they are not as likely to be misinterpreted have an additional value (what was looked for but not found) that may not be utilised.

4.6 Use of standards

Biodiversity information standards have been developed by international initiatives such as TDWG (http://www.tdwg.org). Standards deliver the advantages of providing data in a known format for data exchange and facilitate automation. To be able to use a standard effectively, it needs to be understood. Even Darwin Core, (possibly the simplest of the transport models) has over one hundred and fifty fields. The use of standards is in selecting the appropriate fields to map to rather then trying to fill out the entire standard.

There are detailed schema specifications available for standards but limited documentation on how to use them. It may be considered "background knowledge" in biodiversity data management but as the ALA develops re-usable systems to support more of a citizen science community this type of documentation should be made available. Interfaces to data entry and mapping tools need to be designed with both a new and expert user in mind.

Standards also document controlled vocabularies. Without standard terms and scales it is not possible to use data from different sources in the same analysis. However mapping terms to a common vocabulary is an intensive task requiring domain specific knowledge

4.7 Accessible systems

The processes and systems developed by the Atlas should be publicly available and not only resident on ALA infrastructure. All validation algorithms and tools as well as documentation and process guidelines should be available for others to use or implement on their own infrastructure. The goal is the collection and management of quality information and the best place to achieve this is at the source. It is only when resources are not available at the source that processes should be carried out on ALA infrastructure.

A caveat with externally accessible systems is that they will be modified by users so transparency of modifications will need to be maintained.

4.8 Embedding quality in management process

With the support of collecting organisations, the Atlas should encourage embedding quality considerations into the management process as well as the tools and processes. Technology can enforce certain rules in data collection including mandatory fields and controlled vocabularies however these can always be circumvented. The inclusion of quality considerations in the management of collecting events as well as the detail of the events themselves provides an assurance that information is being collected at the right stage of the process when it most accurate. A key example of this the requirement for a metadata profile of a collecting event (e.g. a survey – methodology, collector(s), identification references, dates, geospatial tools) to be completed and submitted for approval before the event is approved. These profiles can be reused.

It is likely that much of this already takes place but is not associated with the records generated during the collecting event. If technology or process guidelines developed collaboratively would assist in this then the Atlas should look to facilitate.

5. Definitions

- **Precision –** level of detail to which a value is reported
  - E.g. Identification to: genus, species or subspecies,
  - E.g. Decimal coordinates recorded to: 3 decimal places or 1 decimal place

- **Accuracy –** how close is it to the true value
  - E.g. Identification is: certain, uncertain, doubtful
  - E.g. coordinates are within: 10km, 10m of the actual occurrence

- **Consistency** -
  - semantic – same fields have the same meaning
  - structural – same data has the same meaning and is recorded the same way

- **Validation –** checking for errors (see a list of example validation checks below)

- **Cleaning –** report and repair gaps and errors

6. Example validation checks

6.1 Technical

Relatively simple, often able to be automated, checks against the integrity of the data. These may indicate incorrect exports, data mapping, field slippage (e.g. moving 1 column to the right) or data missing at the source

- completeness
  - whether all the data and metadata is available – are all fields present, are all fields filled out
- bounds
  - e.g. days are 1-31 (depending on month)
- data type
  - does Date field contain a date or a number
- data format
  - 01/01/2010 or 01/Jan/10

6.2 Consistency

Application of real world rules to the data. These may indicate incorrect data entry from older records, transcription errors or post processing. Some are complex to implement and require reference data sets to check against. E.g. list of known collectors and collecting habits. These rules can be gathered from data users and analysts.

- taxonomic
  - If identified to species level then there should be a binomial scientific name and entries in genus and species fields
- currency
  - are dates of collection, identification, update and digitisation consistent
- outliers - detect outliers, but not all outliers are errors
  - known species range
  - known environmental range
    - may be a misidentification rather than incorrect coordinates
- geographic
  - are coordinates within identified locality or region
  - terrestrial occurrences in the sea and marine occurrences on land
- collecting patterns
  - does the occurrence detail match the known collecting patterns of the organisation or collector
  - records created after a collector has died (possibly a different collector with a similar name)
  - mammal records attributed to a bird watching group
- Accuracy and precision
  - very high precision or accuracy indicated for records pre GPS (or pre accurate GPS)

- Collecting methods
    - Different survey methods (e.g. transects and area surveys) have particular characteristics, are the records consistent with the method.

7. References

GBIF. 2010. *GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network*, version 1.0. authored by Hill, A. W., Otegui, J., Ariño, A. H., and R. P. Guralnick. 2010. Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. Accessible on-line at http://www.gbif.org. (http://www.gbif.org/communications/news-and-events/showsingle/article/gbif-position-paper-enhancing-fitness-for-use-across-gbif)

Chapman, A. D. 2005. *Principles of Data Quality*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. (http://www2.gbif.org/DataQuality.pdf)

Chapman, A. D. 2005. *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. (http://www2.gbif.org/DataCleaning.pdf)