**Development of a Use Case Library for Biodiversity Data Quality**

# Emily Rose Rees

Supervisor: Miles Nicholls, Atlas of Living Australia

# Abstract

The Atlas of Living Australia (ALA) contains over 60 million records, and over 9 billion records have been downloaded since the Atlas first went online. The quality of data available is thus of high importance, particularly when it is often not being used for the same purposes it was originally collected for and can influence conservation approaches, policies, and future research efforts. As part of an international working group for biodiversity data, a use case library was constructed and analysed to investigate how ALA users determine the quality and fitness for use of records and datasets. Coordinates, scientific name and event date were found to be the most important aspects of a record when determining fitness for use. A series of recommendations are also provided to enhance the quality of the data available through the ALA, improve the user interface and understanding of internal ALA data checks, and improve the use case collection method for future research.

**Key words:** data quality, use case library, biodiveristy data, Atlas of Living Australia

# Introduction

Advances in technology over the past few decades have seen research becoming more global, more complex and increasingly data-intensive. With researchers tackling bigger issues such as climate change and the ongoing biodiversity crisis, they require access to larger amounts of data, both legacy and new (Guralnick et al. 2007). This data needs to be organised, easily accessible, digital and of a high quality. As a result, various online initiatives for data sharing and management have emerged.

Traditional holders of biodiversity data such as museums and herbariums are digitising their collections and making them available to larger databases such as the Global Biodiversity Information Facility (GBIF) and the Atlas of Living Australia (ALA).

The Atlas of Living Australia (ALA) is the Australian Node of the GBIF, and Australia’s national biodiversity database. A collaboration between a number of governmental and national bodies, including CSIRO, Australian museums, and the Departments of Agriculture, Fisheries and Forestry (DAFF), and Sustainability, Environment, Water, Population and Communities (DSEWPaC), the ALA is a collaborative, open infrastructure that pulls together biodiversity data from multiple sources, and focuses on making biodiversity information accessible and usable (ALA 2016).

With so much data available the quality of individual records can vary significantly. In the age of data, data quality is receiving increasing attention, particularly the concept of fitness for use. With such increased attention and importance, it is no wonder these global and local databases are focusing on understanding more about the quality of the records they hold, and looking to improve them where possible.

The ALA is contributing to an international working group on data quality for biodiversity data, a combined initiative of Biodiversity Information Standards (TDWG) and the Global Biodiversity Information Facility (GBIF). The working group is broken down into three task groups:

1. Data quality framework
2. Tools and systems
3. Data quality use cases

At the time this project was initiated, the framework and a series of information collection mechanisms had been developed, including worksheets and google forms. There was a need to engage the international biodiversity analysis and data management community to populate the data.

The collected information was analysed to extract the data elements and report on their use, building a use case library as a resource to research how data is selected for particular purposes. This allows for profiles to be built of the most used information elements, dimensions and criteria, and for these to be fed back into the development of data capture and quality assessment tools.

This project focused on identifying sources of biodiversity use cases by contacting individuals and reviewing documentation and published papers. Potential sources of use cases were communicated with to document the details of the use case. Documented use cases were reviewed to extract individual data quality terms and dimensions. Data quality terms and dimensions were analysed to identify trends or patterns in data quality analysis and compared to existing practices.

# Methods

## Use Case Collection

The use cases for this project were collected using a number of methods to maximise responses.

Based on the 2015 report “Toward A Conceptual Framework for the Assessment and Management of the Fitness for Use of Biodiversity Data” (Veiga et al. in press) a spreadsheet was developed, as well as a simpler google form. Questions included in the google form are listed in the appendix. Lead authors of papers published using ALA data were contacted and asked to contribute, and a number of papers were sent to the ALA Data Quality group that included sufficiently detailed descriptions of fitness for use determination, and how data had been filtered for particular applications. Fitness for use and quality check information from these papers was able to be extracted and transferred across to the use case library.

Three face-to-face interviews with researchers were also conducted. Interviewees were asked to describe their research project, and the checks they apply to the downloaded data before use. Questions asked roughly followed those from the google form, with additional information requested where details were lacking.

## Mapping Data Collected to the Use Case Library

To extract information from the various methods used to collect the data paragraphs, comments, and responses were broken down for mention of data quality criteria. Using the pre-built skeleton for the use case library, these comments were mapped onto relevant criteria, dimensions and information elements. If there was not a matching or equivalent entry in the use case library one was created. Occasionally, criteria statements specific to one use case only were broadened to be more inclusive – i.e. “coordinates must be located within Tasmania” may be broadened to “coordinates must be located within the relevant area of interest”.

The list of criteria also included columns indicating the dimension and information element the criteria was based off, allowing the flow of information to be consistent across the various spreadsheets.

Information regarding the source of the use case, what application the data was being used for, and the owner of the use case was collected to allow for additional analysis, and to request further information should it be required. Use cases were sorted into application groups based on published papers and how the use case had been described. Categories included distribution modelling, database entry, and species list.

## Use Case Library Analysis

From other assessments of data quality and fitness for use studies, most methodologies are based on simple bar graphs to illustrate the frequency with which certain information elements, dimensions and criteria are used (Arnaud et al. 2016). A similar methodology was employed in this study.

The number of use cases associated with each information element, dimension and criteria were counted, as was the number of information elements, dimensions and criteria associated with each use case. Each criteria and dimension was also associated with a fundamental criterion and fundamental dimension respectively, and the number of use cases associated with each was also assessed.

# Results and Discussion

## Methodology and data limitations

This study included data from 26 submitted use cases, as well as two example use cases, which were based on general criteria for studies of that type (ecological niche modelling, ecological gap analysis etc.).

The number of use case submissions suffered from a low response rate. Of the 30 emails sent out to individual researchers, only six use cases were submitted. Few, if any, submissions resulted from general posts on forums including the GBIF Community Site. Greater response rates were had when personally introduced to other researchers at the CSIRO by other members of the ALA.

Often researchers submitting use cases lacked a deep understanding of how their submissions were to be analysed, leading to statements lacking the precision of information required. One researcher, for example, indicated that they primarily used locality text descriptors if they contained sufficient detail to correctly locate the occurrence. Statements such as these are more difficult to transfer to the use case library as they are not based on a clearly defined qualifier, and are more dependent on the researcher, the environment, and the identifiability of the description. Where possible, researchers were asked to provide further information as to what would qualify an entry as sufficient to include in their analysis.

A number of submissions also included information about criteria that wasn’t explicitly exclusionary. One use case in particular included criteria based on an additive scale, whereby not meeting some of the criteria wasn’t grounds for excluding the record, rather that that particular record ended up with a lower score, and was considered of a lower quality. Records that did not meet baseline criteria were, however, excluded. All criteria mentioned in this such manner was recorded and included in the analysis presented here, as even if it was not used on an exclusionary basis, it was still used to determine an aspect of the record’s fitness for use.

Use cases submitted through the google form greatly outnumbered other methods. The google form was designed to be more user friendly than the spreadsheet; as a result, a greater amount of processing of the submitted data is required to transfer the submissions into the use case library.

In general, it was difficult to pull fitness for use assessment criteria from published papers. Papers that did contain a sufficient level of detain regarding such criteria tended to be specifically describing suitability criteria, fitness for use determination and quality assurance or quality control.

A breakdown of the various use case submission methods can be seen in Figure 1.

## Use Case Analysis

The number of information elements, dimensions and criteria used by each use case was counted, and the mean, median and mode determined for each.

For the information elements, the mean, mode and median were all approximately nine (9.214, 9, 9 respectively), with the greatest number of elements associated with a use case being 45. This implies that determining fitness for use of a record or dataset can generally be done based on roughly nine information elements. The number of information elements associated with each use case can be seen in Figure 2.

For the dimensions, the mean, median and mode were 8.36, 5 and 4 respectively, indicating that fewer than ten dimensions are required to adequately assess a record’s fitness for use.

The mean number of criteria per use case was 13.11, the mode 11 and the median 8. When the use case based on 101 criteria was excluded from the mean calculations, the mean was 9.85. This indicated that most use cases used approximately 10 criteria when filtering their data. Figure 3 below shows the number of criteria associated with each use case.

Consistent across these three assessments is that the use case titled “PNT” relies on a significantly larger number of information elements, dimensions, and criteria than other use cases. This particular use case submission stated that any records deemed to be uncertain, doubtful or marked as having ‘fatal’ issues through the internal ALA data quality checks were removed from further analysis. Each of the listed ALA data checks were subsequently inputted into the use case library and associated with this particular use case.

The discrepancy between the PNT use case and others could be due to a number of factors. Failing certain ALA data checks does not necessarily mean a record is unable to be used for some purpose, and so this particular use case could be including criteria that are not crucial to determining fitness for use. As filters and criteria can be applied before downloading the data, it is also possible that researchers only submitted criteria applied post-downloading. If this is the case, a fair amount of data may be missing from this analysis, and this analysis would therefore be focused on post-download processing.

## Quality and Area of Interest

Each of the 257 criteria was classified as either a quality criteria or an area of interest criteria. Criteria such as “precision should be between 0 and 1” and “coordinates must not be generalised” are classified as quality based, whereas criteria such as “record must not be classified as a machine observation” or “coordinates must fall within forest or woodland classed land cover areas” are classified as an area of interest criteria. These classifications were somewhat subjective as they were judged based on the information provided about how the data was used, and not directly indicated by the data user. Of the 257 criteria, 211 were classified as quality based, and 46 as area of interest. This indicates that the majority of criteria used to assess fitness for use are assessing some aspect of the record that is not necessarily specific to the project, and as such could be improved for a majority of applications and uses.

## Criteria

The number of use cases using each criteria was first assessed to determine the most commonly used criteria. As can be seen in Figure 4, there are five criteria at the left end of the scale which have the most number of associated use cases. These criteria include that the coordinates are complete, the scientific name must match a given reference list, the coordinates must be within a given range, the basis of record is well formed and that all the coordinates in a dataset are complete.

Each criteria’s fundamental criterion was also assessed to determine the most commonly used one. As can be seen in Figure 5, DQ measure must be in the range is significantly more popular than any other fundamental criterion. This includes criteria whereby the precision or completeness must be within a given range, where coordinates must be in a range, data must not be generalised and so on. The prevalence of this particular fundamental criterion indicates that given ranges of values, precisions and completeness are the most important aspect of a record to determine its fitness for use.

## Information Elements

With each record based on an information element, assessing the most used elements would indicate where the most effective data quality improvements could be made.

The analysis of the distribution of information elements indicated that there were a number of information elements that had a significantly higher number of use cases associated with them than other information elements. These included coordinates, decimalLatitude and decimalLongitude (of which coordinates are composed), eventDate, and scientificName. The remaining information elements were all used by less than half of the documented use cases. The number of use cases associated with each information element is shown in Figure 6 below.

Coordinates, decimal latitude and decimal longitude all contain location information, and their usage by almost every use case included in the analysis indicates that locality information is consistently one of the most crucial components in determining fitness for use.

When broken down into categories according to the Darwin Core Terms quick reference guide, the prevalence of location information was reinforced. The number of use cases associated with a Darwin Core term in the Location category was almost equal to the number of use cases associated with a term in all the other categories combined. Note that a single use case can appear multiple times in a category if it is associated with multiple terms in that category, i.e. a use case with criteria based on country, locality, and coordinates would be counted three times in the Location category.

The information element analysis clearly showed the importance of locality information, indicating that improving this aspect of a record could have a significant impact on the overall quality of the record.

## Dimensions

Whilst it is clear that locality information is important, to be able to effectively improve the quality of the locality information available an understanding of what aspect of that information is important is required.

As with the information elements, the dimension analysis indicated that there were a number of key dimensions that were used more frequently than other dimensions. Analysing the number of use cases per dimension showed that ‘d:coordinatesValue’, which measured the value of the coordinates, was the most frequently used dimension. Not all use cases were concerned with the value of the coordinates, however, indicating that whilst this is important, it is not the only information used to determine fitness for use. The number of use cases per dimension is shown in Figure 8.

Each dimension was associated with one of nine fundamental dimensions. Value and completeness were the most commonly used fundamental dimensions, indicating that data users are often requiring certain pieces of information as a baseline, and assessing some aspect of the value, be that it lies within a given range, or must be of a given set of values.

Figure 9 shows the distribution of the number of use cases per dimension associated with each fundamental dimension. Again, dimensions value and completeness account for 75% showing their significance.

# Recommendations and Conclusions

The use case library developed highlighted a number of key aspects and attributes of records that researchers were consistently using to assess fitness for use. These aspects were consistent with general understandings of ALA data use.

The recommendations and conclusions drawn from the data are also fairly consistent with studies in the same area (Faith et al. 2013, Arnaud et al. 2016). Used in conjunction with data publishing guidelines, frameworks, and incentives, the quantity and quality of biodiversity data available can be greatly increased.

The majority of the use cases included in this analysis were spatial distribution assessments, and so were using data downloaded from the ALA in a certain manner. It would be worthwhile continuing to collect use cases, particularly of different types, to further understand how biodiversity data is assessed for different applications. A greater variety of use cases, and simply more use cases, will provide better insight into the determination of fitness for use.

## Collection of Use Cases

There was a significant amount of work associated with translating the information received from the google form and the spreadsheet into a format that could be entered into the use case library. On average, the entry of a single use case took one to two hours but some use cases took up to a day or two, particularly when they were based on a higher amount of criteria. In addition, there were generally only about three to four fields in the google form that provided relevant information for this particular analysis. Some responses lacked the detail required for the use case library which made capturing the information they provided difficult. Changing the structure and questions asked on the google form could help address this issue.

A series of check boxes for data users to identify broad areas they filter data based on (location, taxonomy, event information etc.) could be included. This could then open up a series of follow-on questions relating to what aspect of the data was important (completeness, uniqueness, vocab match etc.) as well as which information elements in particular were relevant (coordinates, country, locality etc.). It would also be important for researchers to identify themselves whether each criterion was a quality criterion or an area of interest, as this can be largely dependent on the context and application of the data.

ALA users could be emailed after they have downloaded records and asked to complete the google form. Directly contacting individuals got a better response rate, and so targeting users when they download data means documenting their data filtering process will be at the forefront of their minds throughout the process, making filling out the form easier and more accurate.

Access to a short report on data quality and use case analysis should be provided so that users filling out the google form can better understand how the information they provide will be analysed. This could improve the quality of information provided.

## ALA Data Checks

A number of contributors noted that whilst they were aware the ALA performs data quality checks and makes the results of these available for each record, it was often difficult to correctly interpret those checks. Some contributors requested a “plain English” explanation of the checks, worked examples or additional information about how the check was being performed. If researchers are better able to understand the internal quality checks and reports, it is likely that this feature of the Atlas will be better utilised.

A number of researchers also made comments to the effect that they treated the errors or quality issues present within certain records as being minimal given the large volume of data they were using. Providing a concise explanation of the impacts of data that is not fit for a particular use even though it may otherwise be considered of good quality might help make researchers more aware of the consequences of using records of lower quality.

## Additional Atlas Features

Some contributors indicated that there were some aspects of the Atlas that were lacking, or features that they felt could improve their experience. One contributor mentioned that the ALA lacks models of richness or abundance, that there is no indication as to the expected population in a certain area, or expected species in certain places and so on. Potentially the findings of research papers assessing such things could be incorporated back into the Atlas.

A front end decision tree for filtering data within the ALA was also requested. It was suggested that Atlas users would be able to select which data fields are relevant to them, include worked examples of data quality checks, and indicate how much data each step filters out. Researchers also inquired as to the possibility of then downloading a script for their records, and to run again on different data with the same criterion.

It was also suggested that perhaps the user interface could be tailored for proficiency and intended use. After some period of time, or demonstrated increased proficiency additional features could become available to users, with explanations available as to their functionality.

For users uploading data, providing a drop down selection or autofill based on an authority file (where one exists) would reduce errors associated with nomenclature, misspellings and synonyms (Chapman 2005a, b).

**Acknowledgements**

My sincerest thanks must be extended to all the staff at the Atlas of Living Australia, and in particular to Miles Nicholls, for supervising my work throughout the duration of this project.

Thanks must also go to Alexander Schmidt-Lebuhn for all his efforts coordinating the 2016/2017 CSIRO Summer Vacation Scholars program through which this project was completed.

This project of course would not have been possible without the contributions from researchers detailing how they filtered their data. Thanks to all those who submitted a use case, suggested other researchers, or passed the google form on to someone else.

# Literature Cited

ALA. 2016. About the Atlas.

Arnaud, E., N. Castaneda, J. Ganglo Cossi, D. Endresen, E. Jahanshiri, and Y. Vigouroux. 2016. Final Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity.

Chapman, A. D. 2005a. Principles and methods of data cleaning. GBIF.

Chapman, A. D. 2005b. Principles of data quality. GBIF.

Faith, D., B. Collen, A. Ariño, P. K. P. Koleff, J. Guinotte, J. Kerr, and V. Chavan. 2013. Bridging the biodiversity data gaps: Recommendations to meet users’ data needs. Biodiversity Informatics **8**.

Guralnick, R. P., A. W. Hill, and M. Lane. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters **10**:663-672.

Veiga, A. K., C. Gendreau, A. Chapman, T. Robertson, D. Schigel, and A. M. Saraiva. in press. Toward A Conceptual Framework for the Assessment and Management of the Fitness for Use of Biodiversity Data.

# Figures

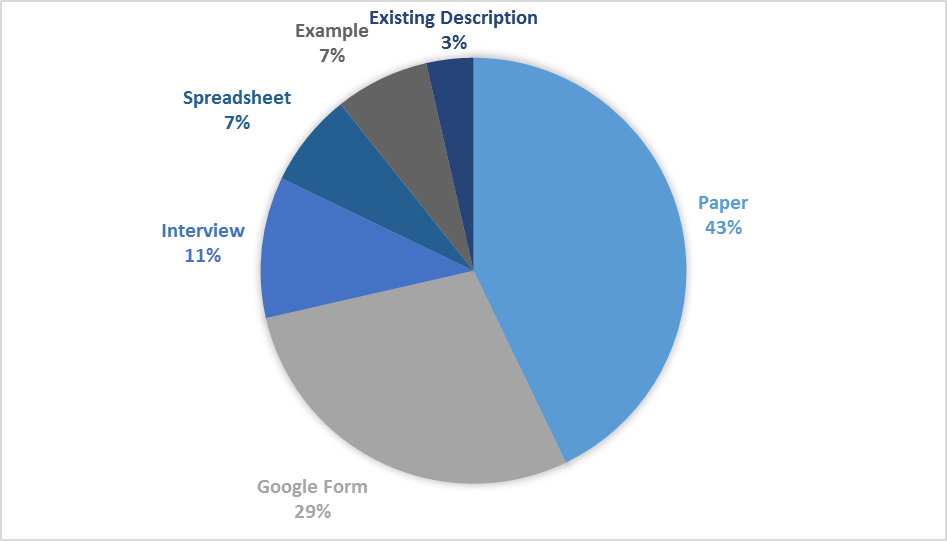


Figure 1 Distribution of Use Case sources

Figure 2 Number of information elements used by each use case

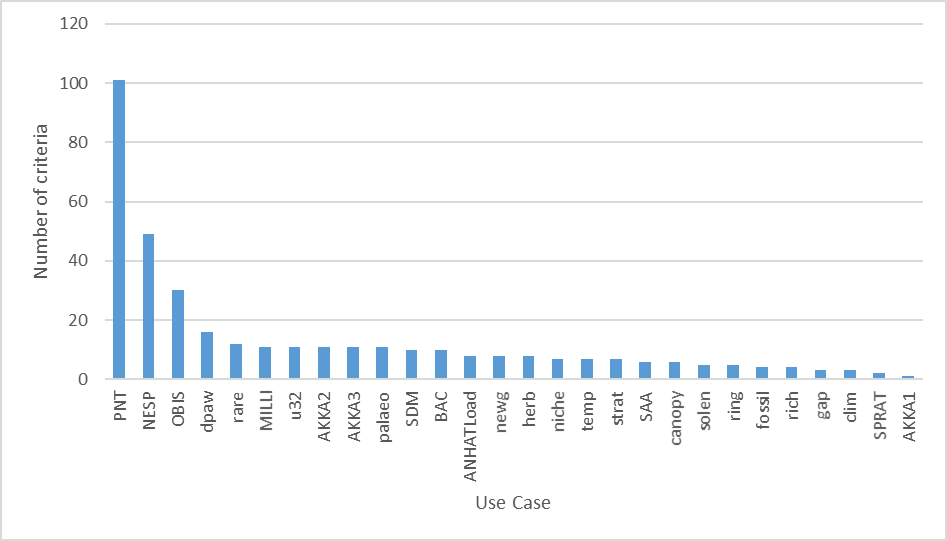


Figure 3 Number of criteria used by each use case

Figure 4 Number of Use Cases associated with each Criteria. Criteria that had fewer than three associated Use Cases are not shown

Figure 5 Number of Use Cases with criteria based on each fundamental criterion

Figure 6 Number of Use Cases associated with each Information Element. Elements that had fewer than three associated Use Cases are not shown

Figure 7 Number of Use Cases associated with information elements in each Darwin Core Terms Quick Reference guide category

Figure 8 Number of Use Cases associated with each Dimension. Dimensions that had fewer than three associated Use Cases are not shown

Figure 9 Distribution of number of use cases per dimension associated with each fundamental dimension

# Appendix

## Google Form Questions

I. Name the data use story

II. Describe the objective of the data use

III. Your Firstname Lastname

IV. Your email

Q1/11. Describe the expected data and information product

Q2/11. Define target audience of the data produce

1. The GBIF network (nodes, head of delegation, GBIF committee members)
2. Data holders (natural history museum, citizen scientists, etc)
3. Biological knowledge experts (individual scientists, academia)
4. Data users (biologists, conservation practicioners, domain experts)
5. Decision makers (government, funding agencies, conventions, organisations)
6. Public stakeholders (educators, private sector, amateurs)

Q3/11. What data sources are needed for the analysis? Name the sources, describe the original requirements, scale, focus and coverage, as well as motivation for use.

Q4/11. Describe the data flow as a set of steps indicating who is performing each step (a scientist, a developer, a system) and what is the purpose of each step.

Q5/11. List the field names as well as quality criteria the data need to meet, and whether the criteria relate to a single record or dataset. Please provide links to any corresponding standards.

Q6/11. Please specify the validation steps that flag data quality issues.

Q7/11. Please specify, if applicable, the steps that enhance your data. How do these improve the quality?

Q8/11. List and link to all tools and systems that are used.

Q9/11. Did GBIF offer suitable data for the analysis? Please specify i) how could the quality and coverage of existing data be improved ii) what additional data fields would you need for the analysis, and iii) how could GBIF better serve your data access and data use needs?

Q10/11. Please identify i) any external, complementary digital data sources that would be valuable for the analysis, ii) holders of non-digital resources that you would like to see digitised, and iii) any technical constraints to data access. Please elaborate and give names and contacts.

Q11/11. Please provide (max 5) URLs, bibliographic references and other comments.