**LIS 545 Term Project Report**

Kari Orth

The Information School, University of Washington

LIS 545: Data Curation

Professor Matthew Mayernik

**Data and Metadata Profile**

|  |  |
| --- | --- |
| **Selected Open-Access Repository:** | <https://datadryad.org/stash> |
| **Dataset URL**: | <https://datadryad.org/stash/dataset/doi:10.5061/dryad.98sf7m0jd> |
| **Dataset: DOI:** | <https://doi.org/10.5061/dryad.98sf7m0jd> |
| **Dataset Name:** | “Humpback whale adult females and calves balance acoustic contact with vocal crypsis during periods of increased separation” |
| **Dataset Contributors:** | Indeck, Katherine; Noad, Michael; Dunlop, Rebecca |

**Data Discussion**:

*What are the data?* The data in this dataset are the movement metrics, call rates, and acoustic levels of humpback whale (*Megaptera novaeangliae*) adult female-calf pairs as they migrate. The data mainly concerns the spatial separation and communication between these parent-offspring pairs.

*Where do the data come from?* According to the “Methods” section of the dataset’s Dryad page, the dataset “was collected using animal-borne recording tags and accompanying boat- and land-based visual observations, as well as recordings from a moored hydrophone array.” The data is derived from the analysis of the movement metrics of tagged adult female-calf pairs during migration.

*Who are the key stakeholders for the data?* There are three key stakeholder groups for this dataset:

* Indeck, Noad, and Dunlop (researchers and data owners)
* The University of Queensland (research institution)
* BG group, BHP Billiton, Chevron, ConocoPhillips, Eni, ExxonMobil, IAGC, Santos, Statoil and Woodside, the United States Bureau of Ocean Energy Management (BOEM), Origin Energy, Beach Energy, AWE, the Holsworth Wildlife Research Endowment & the Ecological Society of Australia (BRAHSS project sponsors)

*Information about data files?* There are four data files available for download through Dryad as well as a README PDF containing additional information about the dataset’s column headings. The data is viewable as XLSX files (Excel spreadsheets).

* ActiveSpace (File 1): This spreadsheet was used to determine the communication active space of adult female-calf vocal signals and whether these pairs modified this active space in response to their behavioral state. The spreadsheet contains two tabs. In the first tab, there are ten total columns. In the second tab, there are five total columns.
* BehaviourModels (File 2): This spreadsheet contains two tabs. The CallRate tab was used to determine if call rate (calls per hour) varied per individual between behavioral states. In this tab, there are seven total columns. The RLMum tab was used to determine if adult females modified their received call levels (dB re 1 µPa) in response to the behavioral state of the pair. In this tab, there are three total columns.
* DepthData (File 3): This spreadsheet was used to examine whether fine-scale dive parameters of the adult females were significantly different between behavioral states. There are seven total columns.
* MovementData (File 4): This spreadsheet was used for the non-hierarchical partitioning clustering method ‘k-means’ to determine behavioral states. There are eleven total columns.

*Does the data set come with any usage restrictions?* This dataset does not appear to have any usage restrictions.

*Is software required to open/analyze the data files?* The only software needed to open and analyze the data files is Microsoft Excel, as the four main files are XLSX files.

**Metadata Discussion**:

*What metadata does the data come with?* This dataset’s publication through Dryad contains key metadata. First, it includes the contributors’ names, their contact information, the publication date, and DOI. Second, it includes keywords assigned by the contributors for findability purposes (e.g. “natural sciences”, “acoustic crypsis”, “behavioural states”, etc.). Third, there is a sidebar menu displaying the dataset’s metrics: how views it has received, how many times it has been downloaded, and how many other publications it has been cited by. Fourth, below the metrics is information regarding the license this work is licensed under. It is licensed under a CC0 1.0 Universal (CC0 1.0) Public Domain Dedication license.

As for the metadata present in the data files themselves: the README PDF includes metadata and additional descriptions for each column in the files. Below is two examples of the data file MovementData’s metadata – first as it appears in the README file and second as it translates to the column headings in the data file:

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

*Where is the metadata?* The metadata can be found through its posting in Dryad as well as in the README PDF and the data files themselves.

*How comprehensive is the metadata provided?* The metadata and descriptions are thorough and helpful for those approaching the data without any prior knowledge. The column headings have accurate names that are fitting and informative of the data contained in them. Further, the descriptions of what the researchers intended these column headings to mean are well-formatted and contain details pertinent to the research data in each column.

*Any metadata standards used to structure the metadata?* Many of the metadata structure suggestions in DataOne’s Data Management Skillbuilding Hub[[1]](#footnote-1) have been implemented in this dataset and the accompanying usage notes. Files have been assigned descriptive files names that “reflect the contents of the file”[[2]](#footnote-2), the research project is described with the appropriate information in Dryad[[3]](#footnote-3), the dataset has been documented and stored using a stable file format[[4]](#footnote-4), and the data that appear in the files match their description in the README PDF metadata.[[5]](#footnote-5)

**Enrichment Suggestions**: To ensure that anyone using this data to replicate the study done in this research project, it would be helpful for either the README PDF or the information accompanying the dataset’s posting on Dryad to include further descriptions of measurement techniques.[[6]](#footnote-6) This could include data collection methods/protocols and instrumentation, as well as notes on the collected data’s “accuracy, precision”, and reproducibility”.

**Publications**: This dataset has been used for a research article ([link](https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.8604)) published to the Wiley Online Library by the same data owners who published the dataset in the open-access repository Dryad. The research article was published on February 9, 2022 as a part of the Ecology and Evolution journal (Vol. 12, Issue 2). The article’s thesis is as follows: “In summary, adult female–calf pairs maintain successful communication during periods of separation by calling more frequently rather than by producing louder calls.”

**Repository Profile**

|  |  |
| --- | --- |
| **Selected Data Repository** | Atlas of Living Australia |
| **Repository URL** | <https://www.ala.org.au/> |
| **Re2data Repository URL** | <https://www.re3data.org/repository/r3d100010918> |

**Reasoning for Repository Choice:** To recap – the data in the selected dataset[[7]](#footnote-7) are the movement metrics, call rates, and acoustic levels of humpback whale (*Megaptera novaeangliae*) adult female-calf pairs as they migrate. The data mainly concerns the spatial separation and communication between these parent-offspring pairs. The data were collected “during the whales’ southward migration away from the breeding grounds, across 4 years off the coast of Peregian Beach, Queensland, Australia.”[[8]](#footnote-8)

This dataset appears to fit well with the data in the Atlas of Living Australia (ALA). According to the description available on Re3data, the ALA “combines and provides scientifically collected data from a wide range of sources such as museums, herbaria, community groups, government departments, individuals and universities. Data records consist of images, literature, molecular DNA data, identification keys, species interaction data, species profile data, nomenclature, source data, conservation indicators, and spatial data.”[[9]](#footnote-9) Further, the ALA “helps to create a more detailed picture of Australia’s biodiversity for scientists, policy makers, environmental planners and land managers, industry and the general public, and enables them to work more efficiently.”[[10]](#footnote-10)The dataset’s content falls within the realm of that available through the ALA, both in terms of geographic location (Australia) and subject matter (biodiversity data).

**Repository Data Submissions:**

*What data will the repository accept?* According to the ALA Submission Guidelines, to submit a dataset “you need to register and have an ALA user account to upload data.” All datasets must adhere to the Darwin Core standard – this is detailed more thoroughly further in the report. The domain is anything that falls under the umbrella of “biodiversity in and around Australia”, and the central purpose of the repository is to store datasets pertaining to records, spatial data, and images of Australian biodiversity. Acceptable data formats and types include:

* Record type
  + Human observation
  + Preserved specimen
  + Observation
  + Occurrence
  + Machine observation
  + Material sample
  + Living specimen
  + Material Citation
* Multimedia
  + Image
  + Sound
  + Video
* Content type
  + GenomicDNA
  + Conservation management
  + Images
  + Pest management
  + Point occurrence data
  + Threats

*What guidance does the repository provide to a potential data submitter as to what should be in the Submission Information Package?* The repository does not provide explicit guidance on what should be in the SIP; however, it is evident from the submission instructions that a submitted dataset must be structured according to the Darwin Core standard. The file(s) can be submitted as Simple Darwin Core, Text, XML, or RDF.

*Does the repository provide any human assistance or consulting to the submitter?* There is a place where a submitter can open a support ticket.[[11]](#footnote-11) It requires the requester’s email and has boxes for a subject and description of the submission issue.

*Does the repository require metadata to be submitted in any specific structure, or according to any specific standard?* The ALA requires all datasets to be in a structured data format, either as a spreadsheet or database, “with an identifier for each record and [other information] in individual columns”[[12]](#footnote-12), following the data standard set by the Darwin Core schema.

According to their dataset submission guidelines, “the first stage in submitting a data set is to produce a file suitable for loading into the ALA.” If the dataset is already in a spreadsheet, the guidelines suggest that the submitter create a copy with Darwin Core column headings or copy the dataset into five data structure templates. They supply these data templates (.xls files) that outline the data structure for a) occurrence data (available as a simple or high quality template), b) species lists, c) multimedia content, and d) species lists.

The Darwin Core standard is “maintained by the Darwin Core Maintenance Interest Group” and “includes a glossary of terms… intended to facilitate the sharing of information about biological diversity by providing identifiers, labels, and definitions. [It] is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information.”[[13]](#footnote-13) Recommended terms are outlined in the quick reference guide (<https://dwc.tdwg.org/terms/>) as well as in their list of Darwin Core terms (<https://dwc.tdwg.org/list/>).

The minimum mandatory fields required by the ALA, all data records must have:

* A unique identifier for each record
* The basis of record (observation, specimen, still image, movie, sound)
* A recognizable species name (scientific or common name)
* an indication of location (coordinates, a recognizable place name, a polygon).

**Data Access Mechanisms:**

*Is a login required to download data? If so, what is the process for creating a login?* There is no login required to download datasets, as the ALA is a “collaborative, digital, open infrastructure.” Creating a login and a profile appears to only be necessary for submitting datasets, not downloading them. To create the login, a submitter would supply the following information:

* First and last name
* Email address and confirmation
* Password
* What their primary affiliation is (government, university, private user, medical research institution, etc.)
* What the name of their organization is
* Country, state/province, and city

After creating an account, an email is sent to the email they supplied to activate their account, and they now have access to their ALA profile.

*Is more than one access mechanism provided?* It appears that the only access mechanism is direct file download. There is a link under the subheading “Digitized Records” in the metadata section of a data resource that says “Download Darwin Core Archive of all Records”, and here, the data files can be downloaded.

*Does the repository display metadata using any specific metadata standard?* As previously mentioned, the ALA requires all datasets to be structured and viewable according to the Darwin Core standard.

*What is included in this repository’s Dissemination Information Package?* While this repository does not include a DIP explicitly named as such, the files are downloadable as Darwin Core records, which appears initially as a zip file. This file typically contains a file titled “eml” (dataset) and another titled “meta” (metadata), both of which are HTML files, and may also contain videos and/or images.

**Additional Information**

|  |  |
| --- | --- |
| **Recommended Data Citation** | *The following recommended data citation follows the precedent set by the citations of other records in the ALA dataset repository.*  Indeck, K., Noard, M., and Dunlop, R. (2023) Humpback whale adult females and calves balance acoustic contact with vocal crypsis during periods of increased separation. <https://doi.org/10.5061/dryad.98sf7m0jd> |
| **Long-Term Preservation**  **Considerations** | The file format that the data and metadata files are in is an important consideration for long-term preservation of research data. Using proprietary formats opens the data up to becoming obsolete and if this happens the files could become unviewable and inaccessible. The data files of this project are currently in XLSX format and the metadata files are currently in XML format. Both of these are widely suggested for file preservation. The XLSX files could be converted to CSV format, as CSV files are considered to be at a lower risk of obsoletion than XLSX.  As for software considerations, there is no specific software needed to open any of the data or metadata files. |
| **Copyright License Statement** | An appropriate copyright license for this dataset would be Attribution 4.0 International (CC BY 4.0)[[14]](#footnote-14). A summary of the license is as follows:   * The dataset can be shared, copied, and redistributed in any medium or format. * The dataset can be adapted, remixed, transformed, and built for any purpose, including commercial. * The dataset must be appropriately attributed by the giving of credit, provision of a link to the license, and indication if any changes were made. * No additional restrictions, legal terms, or technological measures that legally restrict others from anything the license permits may be applied to the dataset. |
| **Human Subject Considerations** | There is no personally identifiable information in the dataset; all data in the dataset pertain to the behavioral states of humpback whale (*Megaptera novaeangliae*) adult female-calf pairs as they migrate. No steps need to be taken regarding data anonymization for privacy or ethical reasons. |

1. <https://dataoneorg.github.io/Education/bestpractices/> [↑](#footnote-ref-1)
2. <https://dataoneorg.github.io/Education/bestpractices/assign-descriptive-file> [↑](#footnote-ref-2)
3. <https://dataoneorg.github.io/Education/bestpractices/describe-the-research> [↑](#footnote-ref-3)
4. <https://dataoneorg.github.io/Education/bestpractices/document-and-store> [↑](#footnote-ref-4)
5. <https://dataoneorg.github.io/Education/bestpractices/confirm-a-match> [↑](#footnote-ref-5)
6. <https://dataoneorg.github.io/Education/bestpractices/describe-measurement-techniques> [↑](#footnote-ref-6)
7. <https://datadryad.org/stash/dataset/doi:10.5061/dryad.98sf7m0jd> [↑](#footnote-ref-7)
8. <https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.8604> [↑](#footnote-ref-8)
9. <https://www.re3data.org/repository/r3d100010918> [↑](#footnote-ref-9)
10. <https://www.ala.org.au/about-ala/> [↑](#footnote-ref-10)
11. <https://support.ala.org.au/support/tickets/new> [↑](#footnote-ref-11)
12. <https://support.ala.org.au/support/solutions/articles/6000195493-how-to-submit-a-data-set> [↑](#footnote-ref-12)
13. <https://dwc.tdwg.org/> [↑](#footnote-ref-13)
14. <https://creativecommons.org/licenses/by/4.0/> [↑](#footnote-ref-14)