University of Tasmania
GPO Box 367
Hobart 7001
Australia

imos.org.au

April 2026

IMOS NESP 5.9

Seabird Aggregated Data Product

Product Technical Document

Version 1.1

Thomas Galindo
Australian Ocean Data Network AODN / IMOS

Australia's Integrated Marine Observing System (IMOS) is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). It is operated by a consortium of institutions as an unincorporated joint venture, with the University of Tasmania as Lead Agent.

IMOS acknowledges the Traditional Custodians and Elders of the land and sea on which we work and observe, and recognise them as Australia's first marine scientists and carers of sea Country. We pay our respects to Aboriginal and Torres Strait Islander peoples past and present.

1. Version History

Version	Date	Comments	Author
v1.0	December 2025	Initial Release	Galindo, T
v1.1	April 2026	Minor Revision	Galindo, T

Citation

Galindo, T. (2026). Seabird Aggregated Data Product Technical Document. Version 1.1. Integrated Marine Observing System.

Copyright/Creative Commons Licence

Table of Contents

2. Background

There are many seabird datasets available in the OBIS Australia Node.

Like many biological datasets, Australian Seabird datasets are disparate.

2.1 Overview

Basis of the Seabird Aggregated Data Product

The goal of the aggregated data product is to flatten the OBIS Australia Node seabird datasets into a single aggregated data product ready for scientific analysis.

3. Methodology

The Seabird Aggregated Data Product (Data Product) relies on the standard Extract, Transform, and Load (ETL) methodology.

3.1 Extract

The Seabird Aggregated Data Product is comprised of the datasets documented here.

Data Source

The underlying datasets are sourced from the OBIS Australia Node.

Additionally, the relevant WoRMS and life stage vocabularies were extracted from theWoRMS marine species website and GBIF vocabularies website, respectively.

Extraction

Darwin core files are archive files containing:

1. An `eml.xml`, a metadata file
2. A `meta.xml`, an explanation file
3. One or more core tables; `event`, `occurrence`
4. Optional tables; `measurementOrFact`, `extendedMeasurementOrFact`

See more details here.

Darwin Core to Parquet Translation

The original darwin core data is validated against the following schemas before serialisation to intermediary parquet files:

Occurrence

name	type	nullable
id	string	true
modified	string	true
bibliographicCitation	string	true
references	string	true
institutionCode	string	true
collectionCode	string	true
basisOfRecord	string	true
occurrenceID	string	**false**

catalogNumber	string	true
recordNumber	string	true
recordedBy	string	true
individualCount	int64	true
organismQuantity	string	true
organismQuantityType	string	true
sex	string	true
lifeStage	string	true
occurrenceStatus	string	true
associatedOccurrences	string	true
associatedReferences	string	true
otherCatalogNumbers	string	true
occurrenceRemarks	string	true
organismID	string	true
preparations	string	true
materialSampleID	string	true
eventID	string	true
fieldNumber	string	true
eventDate	string	**false**
eventTime	string	true
year	string	true
month	string	true
verbatimIdentification	string	true
verbatimEventDate	string	true
samplingProtocol	string	true
sampleSizeValue	string	true
sampleSizeUnit	string	true
waterBody	string	true
country	string	true
stateProvince	string	true
county	string	true
locality	string	true

minimumDepthInMeters	int64	true
maximumDepthInMeters	int64	true
decimalLatitude	float64	true
decimalLongitude	float64	true
coordinateUncertaintyInMeters	int64	true
coordinatePrecision	float64	true
footprintWKT	string	true
identificationQualifier	string	true
typeStatus	string	true
identifiedBy	string	true
dateIdentified	string	true
identificationReferences	string	true
identificationRemarks	string	true
scientificNameID	string	**false**
taxonConceptID	string	true
scientificName	string	**false**
kingdom	string	true
phylum	string	true
class	string	true
family	string	true
genus	string	true
subgenus	string	true
specificEpithet	string	true
infraspecificEpithet	string	true
taxonRank	string	true
scientificNameAuthorship	string	true
vernacularName	string	true

Note that rows missing `decimalLatitude` *or* `decimalLongitude` *are dropped*

ExtendedMeasurementOrFact

name	type	nullable

eventID	string	true
occurrenceID	string	true
habitat	string	true
sampleSizeUnit	string	true
sampleSizeValue	string	true
samplingEffort	string	true
samplingProtocol	string	true
behavior	string	true
individualCount	string	true
lifeStage	string	true
organismQuantity	string	true
organismQuantityType	string	true
sex	string	true
type	string	true
measurementAccuracy	string	true
measurementDeterminedBy	string	true
measurementDeterminedDate	string	true
measurementID	string	true
measurementMethod	string	true
measurementRemarks	string	true
measurementType	string	true
measurementTypeID	string	true
measurementUnit	string	true
measurementUnitID	string	true
measurementValue	string	true
measurementValueID	string	true

Note that rows missing both eventID and occurrenceID are dropped

Survey Type

Additionally the `surveyType` column is added.

This external context indicates if the survey is a `Tracking` survey or an `At-sea observations` survey.

name	type	nullable
surveyType	string	false

Validation

Being sourced from an OBIS Node, the data is of high quality and subject to stringent dataset structures.

This goes a long way to simplifying the extraction procedure.

That being said, we conduct the following validation to ensure data quality.

Multi-Format Timestamp Normalisation

The extract stage checks that `eventDate` values are present (non-null and non-empty); the recommended format is ISO 8601 (`^\d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}Z$` , e.g. `2012-11-08T12:31:00Z`).

During transform, `eventDate` values are parsed to dates using a coalesced multi-format strategy that accommodates 11 distinct patterns found in real-world data, including:

- Standard ISO 8601 with UTC marker (`%Y-%m-%dT%H:%M:%SZ`)
- Space-separated datetime (`%Y-%m-%d %H:%M:%S`)
- Date only (`%Y-%m-%d`)
- With minutes only (`%Y-%m-%dT%H:%M`)
- With milliseconds (`%Y-%m-%dT%H:%M:%S.%fZ`)
- With nanoseconds
- Both with and without the `T` separator
- Both with and without a UTC timezone marker

Non-conforming values that do not match any pattern are set to null.

Scientific Name and ID Match

The `verbatimIdentification` , `scientificName` and `scientificNameID` darwin core columns are used in conjunction to ensure the identified species is interoperable between datasets.

We ensure the `scientificName` is up to date with the `scientificNameId` data found in the World Register of Marine Speices (WoRMS).

Metadata

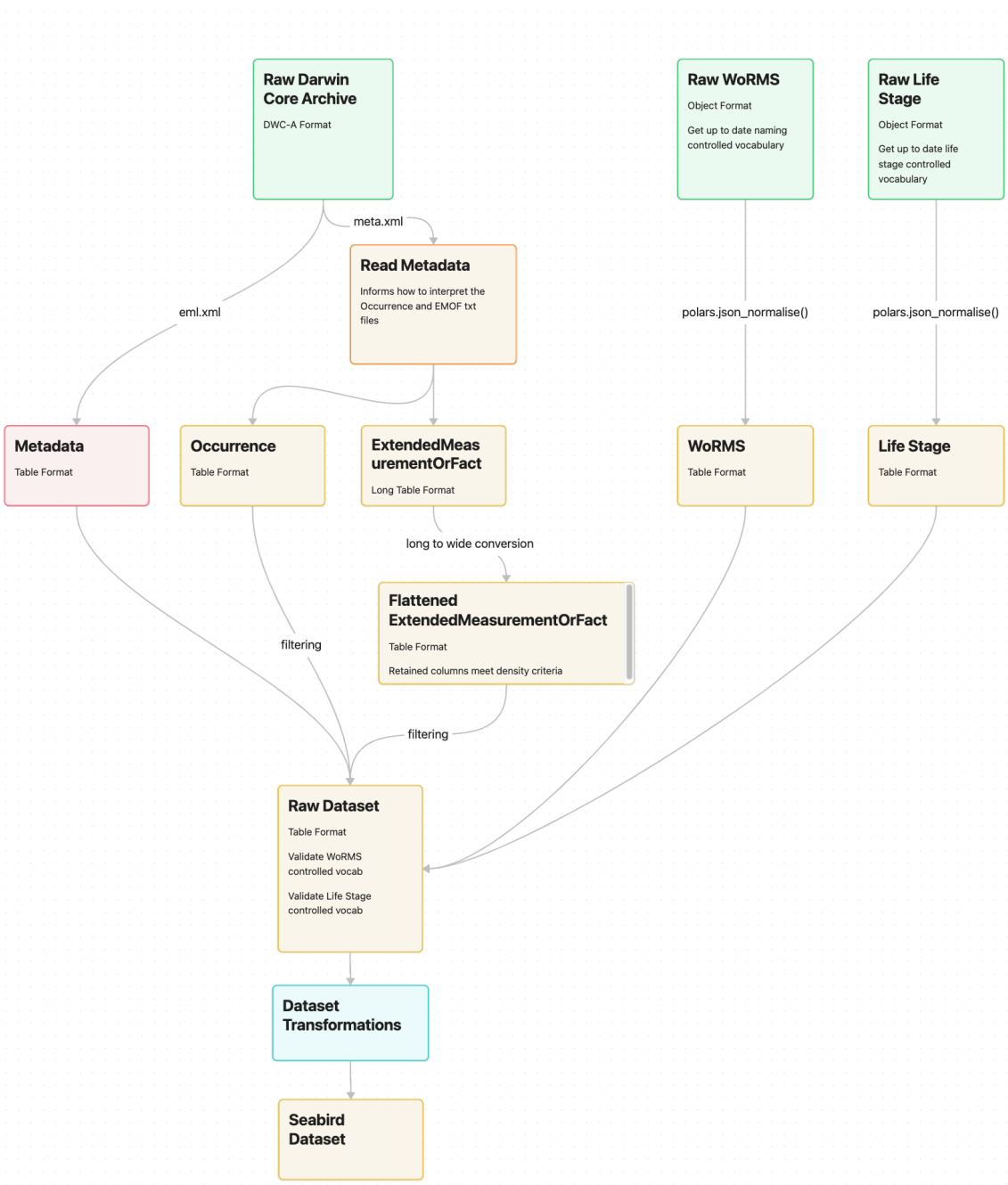
Darwin Core archives must have a `eml.xml` file, which contains traditional dataset metadata.

We pull the following from the `eml.xml`:

name	description
dataset_name	the dataset long name
start	the dataset start date
end	the dataset end date
geographic_bounds	the geographic bounding coordinates
contacts	the dataset conctacts
abstract	the dataset abstract
pub_date	the dataset publication date
version	the dataset publication version

3.2 Transform

The transform step synthesises the darwin core, WoRMS and GBIF Life Stages tables:



Aggregation

The individual Darwin Core archive parquet are aggregated vertically:

Pre-Filtering

Pre filtering steps are carried out to ensure valid transformations.

Drop incomplete extended measurement or fact

Valid extended measurement or fact tables must point to either an occurrence or event row. Rows pointing to neither are dropped.

Drop all null columns

Columns completely devoid of information are dropped.

Drop missing latitude and longitude

Rows missing `decimalLatitude` or `decimalLongitude` are dropped.

Drop missing quantity

Rows missing `organismQuantity` are dropped.

Extended Measurement or Fact Pivot

The extended measurement or fact table is a long table with a m:1 relation to the event and occurrence tables.

The extended measurement or fact (EMOF) table therefore must be pivoted to a wide format in order to produce a single parquet output. This involves a trade off between table width and data retention. The measurements retained were arbitrarily cut off at 10,000, which is where a value is present for ~10% of the events/occurrences:

measurementType	count
Sea state	27894
Sea surface temperature	26444
Depth (m)	25186
Salinity (psu)	21772
Wind direction (deg)	20768
Air Pressure (hPa)	18892
Wind Speed (knt)	18262
Wind Force (Beaufort scale)	17292
Cloud cover	17229

Atmospheric temperature (deg C)	16944
Cloud cover (oktas)	13744
CUTOFF	**-----**
Visibility	9363
Bird behaviour	3397
ship speed	2557
Wind direction	93
Air temperature	93
Depth	88
ship activity	75
Cetacean behaviour	62
precipitation	42
Culman Height	26
Animal Mass	26
Tag Type	26
Culman Length	26
LifeStage	13
Seal behaviour	8
sun glare	4

Extended Measurement or Fact Value Merging

Many measurements were found. Careful analysis of values allowed merging of the following `measurementType` was valid with minimal value transform:

Wind Speed

`Wind Speed (knt)`

`Wind Force (Beaufort scale)`

Wind Direction

`Wind direction (deg)`

`Wind direction`

Air Temperature

`Air temperature`

`Atmospheric temperature (deg C)`

Depth

`Depth (m)`

`Depth`

As the validity of any value merges is extremely sensitive to the value data itself, the ETL process reports any mutation of the underlying measurement data to AODN for further investigation.

Occurrence Table (Core Table)

The Seabird datasets were found to only contain Occurrence Core tables.

1. Update empty strings to null
2. Join EMOF table
3. Join metadata table
4. Construct and validate all source urls using download link and version
5. Life stage values are validated against the GBIF Life Stage Vocabulary
6. Timestamps are validated and updated to dates
7. WoRMS taxon hierarchy are mapped via the `scientificNameID`
8. Update the output table metadata with the ETL timestamp

Column Renaming

The following columns are renamed from internal names to output column names before final validation:

Internal Name	Output Column Name
`dataset_id`	`datasetName`
`survey_type`	`surveyType`
`Wind Direction`	`windDirection`
`Air Temperature`	`airTemperature`
`Wind Speed`	`windSpeed`

`Depth`	`depth`
`Air Pressure`	`airPressure`
`Sea State`	`seaState`
`Salinity`	`salinity`
`Cloud Cover`	`cloudCover`

Geospatial Enrichment

The output dataset is enriched with the following geospatial columns:

- `h3Index` - A hexadecimal string representing a H3 polygon at resolution 4, derived from the `decimalLatitude` and `decimalLongitude`
- `australianMarineRegionsTags` - A `|` separated tag column featuring common Australian marine regions

3.3 Load

Finally, the transformed Parquet is loaded to the `processing-stored-bucket` at path `datauplift/seabird/seabird.parquet`.

4. Flow Advantages

Reproducibility the original seabird darwin core files require relational database understanding and specialised software to load and join tables. The ETL procedure makes this process reliable and replicatable

Space Efficiency raw original data is reduced to **~5MB** with modern compression and file formats (Parquet), optimising storage and access

Time Efficiency scheduled updates allow users to access up-to-date aggregated data without re-implementing the aggregation steps and validation

Cloud-native workflows reduce I/O overhead