Atlas of Living Australia
ala.org.au

ALA Extended Data Model Producers' Guide

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

CSIRO

The ALA is made possible by contributions from its partners, is supported by NCRIS and hosted by CSIRO.

ALA Extended Data Model Producers' Guide

2022-08-03

Introduction

The ALA has traditionally collected occurrence records. Records that record the what, where and when of an organism or group of organisms. This data can be used for many purposes but it falls short when researchers want to identify records that have been collected according to a consistent protocol or which are related by, for example, being part of a survey or collecting expedition.

The ALA Extended Data Model project is designed to capture more structured data surrounding the sort of activities that went into getting the occurrences, how the information about the occurrence is derived and supporting information. To begin with, we are going to accept data that is structured around the activities that people (or machines) undertook to collect occurrences and any measurements or other information that is pertinent.

This document is a guide for those who wish to create Event Core Darwin Core Archives that can be accepted by the Atlas of Living Australia.

1.1 Preferred Practices

Scattered throughout the document are lists of "preferred practices." We're not calling them "best practices" yet but they represent what will give the smoothest, most consistent route for your data to travel from your database (or spreadsheet), to ALA ingestion, through the ALA processing pipelines, into the ALA and out to many eager data consumers.
One repetitive element of these practices is the handling of vocabularies. Ideally, anything other than descriptive text should be drawn from a standard vocabulary. The most useful vocabularies are linked data, drawn from the semantic web and are represented by URIs that can be resolved externally, so that both humans and machines can go to a reliable source of information about what something means.
Internal identifiers, such as **eventID**, **occurrenceID**, **measurementID** and the like can be anything that you choose. Just make sure that they are unique. [1]When constructing identifiers from a database or other data source, section 3.2.3 has some strategies that have worked reasonably well.
Identifiers that refer to an external definition, such as **measurementTypeID** should, ideally be represented by resolvable URIs. This isn't always possible, but the order of preference should be:

1. A URI that links to a public, well-known and well-used vocabulary. For example, http://qudt.org/vocab/constant/StandardAtmosphere http://www.opengis.net/def/uom/OGC/1.0/degree or https://creativecommons.org/licenses/by/4.0/
2. A URI that links to a public vocabulary that you have constructed. Research Vocabularies Australia[1] provides tools to allow you to build your own vocabulary. As a general guide, follow the practices of linked data [2] and Cool URIs [3] when constructing your own vocabulary.
3. A non-resolvable URI. A full URI is still useful because it provides an unambiguous namespace when attempting to figure out what concept is being used.
4. A term drawn from a readily available, published vocabulary
5. Something you made up on the spot

Terms that are not explicitly identifiers (with the ID on the end of the term) should still draw from a consistent vocabulary. You may still choose to use a URI. If the vocabulary uses URIs but you don't want the clutter of using them, then use the simple name part of the URI and drop the name space. For example, if the term is http://purl.org/dc/terms/created then use created as the vocabulary term. This approach is often used when describing the terms present in a Darwin Core Archive (see section 0).

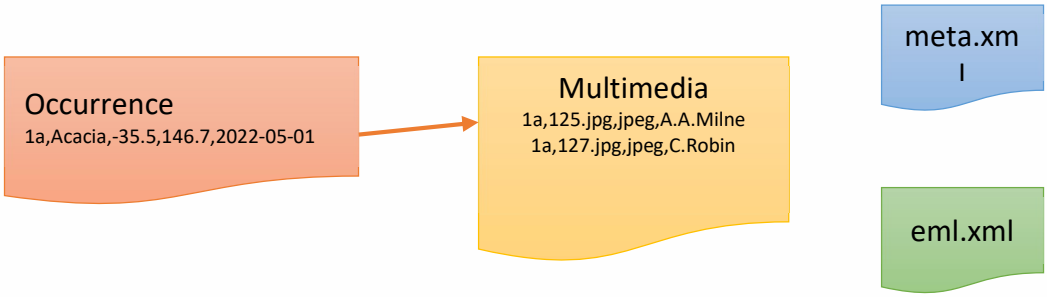
[1] https://vocabs.ardc.edu.au/

2 Building a Darwin Core Archive

The ALA accepts datasets in the form of Darwin Core Archives [4] (DwCA). A DwCA is a package of related tabular data, along with a schema describing the structure and meaning of the data and a metadata statement.

2.1 Structure of a Darwin Core Archive

The basic structure of a DwCA is a core table (a comma- or tab-separated file) and a number of extension tables. These tables are linked by a core id, a single identifier that links a single record in the core table with zero or more records in each extension. When a DwCA is processed, each core/extension record bundle is handled as a single unit.
An example DwCA is shown below, where the core table contains occurrences and there is a multimedia extension.



In this example, the core occurrence table contains the traditional what-where-when of an ALA occurrence record. The multimedia extension table contains information about the images that are associated with the occurrence. The records are linked by an identifier – 1a in this case. The result in this example is a single occurrence record with scientific name Acacia a latitude of -35.5, a longitude of 146.7 and a date of 2022-05-01. The occurrence record is attached to a multimedia extension with two records. One has an image identifier of 125.jpg a format of jpeg and a rights holder of A.A.Milne. The other has an image identifier of 127.jpg a format of jpeg and a rights holder of C.Robin.
Tables are comma- or tab-separated value files. There is a certain amount of latitude allowed in how these files are formatted, with options described in the schema file, described below.

2.1.1 The schema file

The meta.xml file contains a schema description for the DwCA. It is an XML document with a root element of archive. A typical meta.xml file looks like:

```
<?xml version="1.0"?>
<archive xmlns=http://rs.tdwg.org/dwc/text/ metadata="eml.xml">
  <core encoding="UTF-8" linesTerminatedBy="\n" fieldsTerminatedBy="," fieldsEnclosedBy="&quot;"
ignoreHeaderLines="1" rowType="http://rs.tdwg.org/dwc/terms/Occurrence">
    <files>
      <location>occurrences.csv</location>
    </files>
    <id index="0"/>
    <field index="0" term="http://rs.tdwg.org/dwc/terms/occurrenceID"/>
    <field index="1" term="http://rs.tdwg.org/dwc/terms/scientificName"/>
    <field index="2" term="http://rs.tdwg.org/dwc/terms/decimalLatitude"/>
    <field index="3" term="http://rs.tdwg.org/dwc/terms/eventDate"/>
```

```
  </core>
  <extension encoding="UTF-8" linesTerminatedBy="\n" fieldsTerminatedBy="," 
fieldsEnclosedBy="&amp;quot;" ignoreHeaderLines="1" rowType="http://rs.gbif.org/terms/1.0/Multimedia">
    <files>
      <location>multimedia.csv</location>
    </files>
      <coreid index="0"/>
      <field index="1" term="http://purl.org/dc/terms/identifier"/>
      <field index="2" term="http://purl.org/dc/terms/format"/>
      <field index="3" term="http://purl.org/dc/terms/rightsHolder"/>
  </extension>
</archive>
```

Within the archive element there is one core element. This describes the core table. The attributes give the type of data in the table and the CSV encoding to use when reading the table:

- **encoding** The character encoding for the file.
- **linesTerminatedBy** The character sequence at the end of the record (line). Newline and carriage return can be represented as the sequences \n and \r respectively.
- **fieldsTerminatedBy** The character that separates fields in the record. A tab can be represented by \t.
- **fieldsEnclosedBy** The character used to enclose fields that have line or field terminators in them. Generally, in CSV fields, this is the double quote character ("). Since this character has a special meaning in XML, it is escaped as the sequence " The meta file does not allow for a separate escape sequence, such as \. Instead, enclosing characters are escaped by repeating them. Eg. to include *Adele sp. "Floria" (Nurke, 2022)* in the CSV file, you would enclose it as "Adele sp. ""Floria"" (Nurke, 2022)" rather than "Adele sp. \"Floria\" (Nurke, 2022)" (Note that the vertical double quote " is used, rather than the open/close quote pair "".)
- **ignoreHeaderLines** The number of lines at the start of the file that provide header information. These lines are skipped when reading the data.
- **rowType** The URI of the class of record that the file contains.

The files element contains a number (almost always one) of location elements that give the relative location of the file that contains the data.
The id element gives the index (counting from zero) of the field that contains the unique identifier for each row.
The field element lists the fields that are to be loaded for each record. Each field element contains an index attribute that gives the position of the field in the file and a term attribute that gives the URI of the term for that field. Terms can either be proper URIs or simple names, if there is no published URI.
An extension element is similar to the core element, except that it contains a coreid element, rather than an id element, giving the location of the foreign key that links the extension records back to the core record.

2.1.2 The metadata file

The eml.xml file contains a document in the Ecological Metadata Language (EML) [5] See 2.1.4 for more information.

2.1.3 Preferred Practices

The key preferred practice for constructing DwCAs for the ALA is that you should assume that, at some point, someone is going to have to manually open up your archive and examine it. The

conventions we recommend tend to make produced DwCAs human-friendly as well as machine-friendly.
The other preferred practice is to ensure that your data is compatible with the most common CSV reading and writing libraries.

- Use a tool to generate the meta.xml file. Hand-building the file gets old really quickly and leads to a lot of mistakes.
    - https://github.com/AtlasOfLivingAustralia/event-core-tools is designed to work in concert with this document.
    - Keep the field elements in index order
- Use UTF-8 as a file encoding. UTF-8 ensures that any accented characters, for example author names, are consistently handled.
    - Beware of assuming that a source file is encoded in UTF-8; Windows and OSX have a bad habit of sneakily saving files in other encodings or assuming that the file is in another encoding. In particular, opening a CSV in Excel may have the program assume that it is encoded in the default system encoding (Sometimes Mac OS Roman in the case of OSX or Windows-1252 in the case of Windows). Similarly, saving the file may mess around with the encoding. If you are using OSX, the command file -I {filename} will provide a guesstimate as to the file encoding. If you are using linux, the equivalent command is, annoyingly, file -i {filename} Similarly, if you find yourself encodingly embarrassed, iconv -f {original_charset} -t {new_charset} {originalfile} > {newfile} will convert the file.
- Avoid the UTF byte order mark [6] at the start of the file. This is a special character sequence that marks the file as being encoded in UTF-8. Many CSV libraries will not recognise the sequence, leading to errors that will have you believing in conspiracy theories, since the cause is often made invisible. The unix utility **od** will allow you to have a look at the start of the file, via od -a {file}. If you have a byte order mark, dos2unix {file} will remove it; it will also replace a carriage-return/newline line ending with just a newline.
- Ensure that there is a header row with the names of the terms in each column. This will help greatly when you pull the file into a spreadsheet. Generally, the last element of the term URI is sufficient, eg. use scientificName for http://rs.tdwg.org/dwc/terms/scientificName
- Keep the files consistent in terms of CSV dialect.
    - Use the same line terminator in each file, either "\n" for just a newline (unix) or "\r\n" for a carriage-return/newline (Windows)
    - Either tabs or commas are acceptable. Tabs are harder for humans to read, since it is hard to distinguish between a tab, a space or multiple tabs. Commas will have you enclosing data of the form "Agrilinae Gory & Laporte, 1835" a lot.
    - Use double quotes for enclosures. This must be expressed in XML terms as "
- The preferred approach is a to have something meaningful as the core and extension identifiers - something like eventID or occurrenceID. This aids people attempting to interpret the files. However, this is not a requirement.
    - If you don't have a common term, then use a made up value, either a UUID or an integer count. In this case, set the id and coreid elements to have an index of 0 and omit the index-0 field element.

2.1.4 Metadata

A good Darwin Core Archive should contain a metadata statement in the eml.xml file describing the data in the archive, its source, licencing conditions and any other relevant information. The structure of an eml file is available online.[2] It is recommended that the metadata should contain at least:

- An **alternativeIdentifier** that allows the data to be located in the source system that provided the data. This identifier helps identify the source of the data.
- A **title**
- A **creator** with enough information to allow a user of the data to contact the originator of the data or their organisation. An **individualName** or **organizationName** and an **electronicMailAddress** would suffice.
- A **pubDate** giving the publication date of the data.
- An **abstract** with a description of the data. The EML standard contains other elements that can be used to further describe the data in detail but a simple, human-readable description makes the data useful.
- An **intellectualRights** section containing the copyright holder or other rights holder.
- A **licensed** element that gives the licence under which the data may be distributed. The ALA recommends using the Creative Commons Attribution licence (CC-BY 4.0[3]) if possible.
- If available, a **distribution** element giving a source for the data.

2.2 Limitations of Darwin Core Archives

A DwCA looks suspiciously like some sort of relational database in a parcel. It is not. It represents a "star" schema with a central core and surrounding extensions. This means two things that are going to be very relevant to extended data:

- records are always tied by a single core identifier and records from one extension that are tied to those from another extension still have to be accessed via the core identifier;
- many to one relationships, of the sort where you have a reference table of site information, are not allowed.

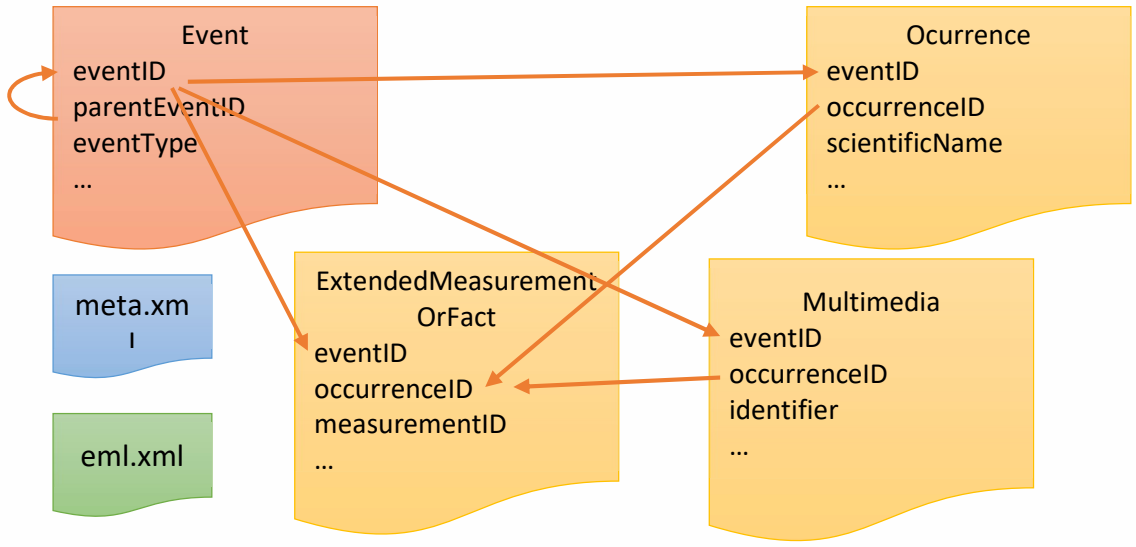
3 The Event Core Archive

An event core archive is an archive that puts the activities used to gather information at the centre of the data. A typical event core archive models some sort of regular survey activity where someone returns to a specific site several times, takes measurements and samples, records sightings of occurrences and takes photos. All the information is tied together by the actions taken to gather the information.

An event core archive has a core table of events and extension tables of whatever information surrounds those events: occurrences, measurements, photos and anything else that might be pertinent. Events can (but do not have to) form a hierarchy. For example, every sample taken while visiting a site is regarded as a child of that site visit. As another example, the events that follow the collection of a specimen, such as accession or condition reporting, are regarded as children of the collection event.

A typical event core DwCA is structured as follows:

[2] https://eml.ecoinformatics.org/schema/eml_xsd.html#eml
[3] https://creativecommons.org/licenses/by/4.0/



This DwCA contains a list of events. The events are linked together in a parent child hierarchy. Some events have one or more occurrences recorded with them. Both events and occurrences can have a measurement or fact (assertion) linked to them. For example, a site visit may have weather readings and an occurrence may have a wing-span.

The primary table follows the Darwin Core Event class4 and uses eventID as the core identifier. There are several extension tables. The occurrence table holds occurrence information. The extended measurement or fact table contains measurements for both events and occurrences. Similarly, the multimedia table holds images about both events and occurrences.

Due to the limitations of the DwCA format, discussed above, the extensions can only be linked by the eventID identifier. Which records apply to occurrences and which records apply to events is determined by a convention discussed below.

3.1 Extending Record Linkages

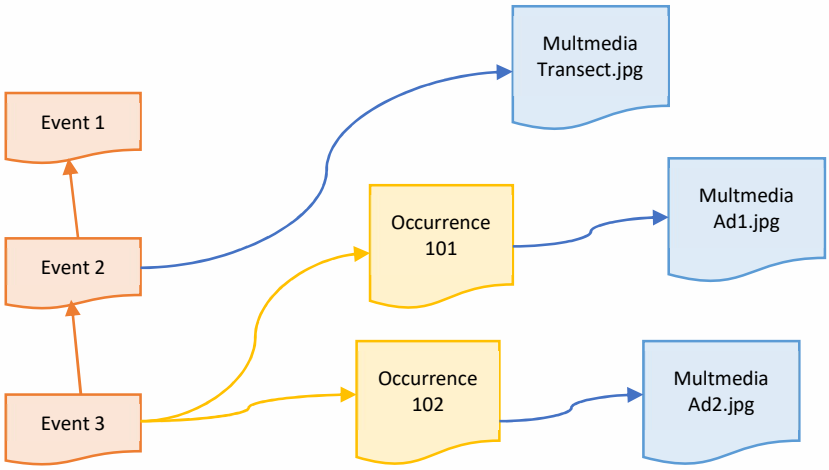
There is a general trick to extending record linkages beyond the simple star format expected by DwCA. This trick comes from the OBIS ENV-DATA model [7]. The core and extension tables are tied together by a common identifier (eventID) in this case. The extension tables contain additional identifiers that can be used to more directly attach the information to an occurrence or other item down the line.

In the example above, occurrences are linked to events. Multimedia entries can either be linked to the events (eg. a photo of the transect) or a specific occurrence (eg. a recording of birdsong). If the eventID is present in the multimedia table but the occurrenceID is empty, then the entry is connected to the event only. If both the eventID and occurrenceID are present, then the entry is connected to the occurrence. As an example, with the following set of tables:

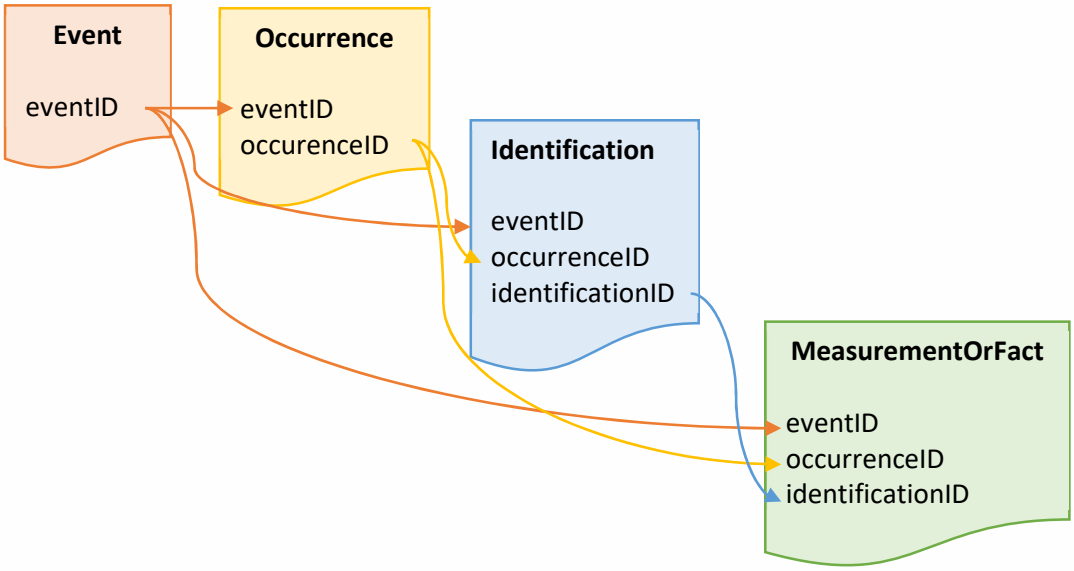
Event			**Occurrence**			**Multimedia**		
eventID	parentEventID	eventType	eventID	occurrenceID	scientificName	eventID	occurrenceID	identifier
1		SiteVisit	3	101	Acacia dealbata	2		Transect.jpg
2	1	Sample	3	102	Acacia dentifera	3	101	Ad1.jpg
3	2	Observation				3	102	Ad2.jpg

[4] https://dwc.tdwg.org/terms/#event

The interpreted structure would then be:



As another example, a collecting expedition may collect a number of specimens. These specimens are later identified, perhaps multiple times, with the identification information recorded as measurement or facts. Everything is connected to events, to model the star schema but there are separate links between the various tables, with a schema as follows:



If a measurement or fact has the eventID, occurrenceID and identificationID all set, then it is associated with the identification. If it only has eventID and occurenceID set then it is associated with the Occurrence and just the eventID associates the measurement with the event itself.

3.1.1 A Note on Implementation

Extended record linkages are a convention and are required to be processed correctly by the software that reads the DwCA. It is up to the processing software to correctly parcel out information based on a common event identifier.

In addition, some linkages are based on something connected to a record linked to a parent event. For example, if there is a collection event that has an occurrence record, there may be child events that are related to the occurrence, such as an accession date, a sequence of identifications or a

series of condition tests. In these cases, the occurrence may not be directly associated with the DwCA star record; this will not stop the data from loading but may require some inference by those using the data.

3.1.2 Event

The event[5] table holds the key information about the activities that led to an occurrence[6]. An event table contains the following essential fields:

- **eventID** A unique identifier for the event.
- **parentEventID** The unique identifier for the parent event, if one exists
- **eventType** A term that gives the type of activity, preferably from a controlled vocabulary.
- **eventDate** The date on which the event took place

Any terms from the DwC event class can also be present, as well as the following non-DwC terms:

- **eventName** A title for the event, eg. "Lake Albert Arthropod Sweep"
- **samplingProtocolID** An identifier for the sampling protocol, for use with published protocols. This will, generally be a link to an external document.

Additionally, the event table may also contain location information (see 3.3 below) and record-level information[7], rather than have a separate location or record table.

3.1.3 Inheritance

The following terms are assumed to inherit from parent event to child event, if null in the child event:

- All record-level information, particularly institution, collection, dataset and rights/licensing information.
- All non-verbatim location information, particularly country, stateProvince, locality, decimalLatitude, decimalLongitud, geodeticDatum and footprintSRS.
- The eventDate and other date/time information.
- The habitat

3.1.4 Preferred Practices

- The DwCA archive standard allows for a local linking identifier to be used for the core and extension identifiers, which is then ignored during ingestion. Unless there is a very good reason for doing so, the eventID should be used as the linking identifier.
- The event type should be drawn from the GBIF event type vocabulary at https://registry.gbif-uat.org/vocabulary/EventType
    - The event type hierarchy implied in the vocabulary is *not* strict.
- The event date should at least be in ISO 8601[8] form, eg. 2022-02-18. If there is a more accurate date/time combination, then a full ISO 8601 date-time should be used including zone offset, eg. 2022-02-18T13:10:15+10:00.
    - If the event date is less accurate than a day, use the ISO 8601 standard for month-level or year-level times, eg. 2022-02 or 2022.
    - If the event date represents a range, use the ISO 8601 standard for representing date ranges. The standard allows a start/end instant to be defined by putting the varying information after a slash. For example, a full range would be 2021-12-

[5] https://dwc.tdwg.org/terms/#event

6 The business of the ALA is still occurrences. All the spinach on top is, essentially, to allow people to find and interpret occurrences in a way that makes sense.

7 https://dwc.tdwg.org/terms/#record-level

[8] https://en.wikipedia.org/wiki/ISO_8601

28T10:00:00+10:00/2022-03-21T16:08:07Z something that is over two hours by 2022-03-07T08:30:00Z/10:30:00Z, something over two days by 2022-03-07/08 something over 32 days by 2022-03-07/04-08, something over two years by 2021/2022.
    - Keep precision within the sensible bounds of the span. A multi-day event can quite sensibly be recorded as 2022-03-04/06 rather than 2022-03-04T10:41:12+10:00/2022-03-06T15:28:31+10:00
    - eventTime, day, month, year, startDayOfYear and endDayOfYear are deprecated.
- Events that have a true nested-doll structure (eg. a survey covers a series of site visits which each over a series of observations) should have date ranges that encompass the child events. For example, two observations at 2021-08-12T10:00:00Z and 2021-08-12:15:30:00Z should fit into the site visit at 2021-08-12T09:00:00Z/16:00:00Z which should fit into the survey range of 2021-07-01/2022-06-30.
- Similarly, locations within a nested-doll structure should have parent locations encompass the child locations.
- If a single dataset contains multiple expeditions or other top-level activities, use the eventName to title the top-level activity. This situation is quite common in data feeds from the databases of museums, government bodies and other large institutions.
- The ideal form of the samplingProtocolID is the DOI of the published protocol or the URI of a resolvable semantic resource.

3.1.5 Identifier Construction

There's a tendency for data sources to have a single identifier for many aspects of an activity, that will get broken down when placed in an event core archive. If new identifiers need to be constructed and there is a clear structure to follow, then, identifiers can be built by extending the base event identifier. If the event identifier is a URL, then child events can be identified by adding further elements to the path. For example:

- A site visit identifier is https://id.test.com/survery/A549/visit/2022-01-14
- The first sample can then be https://id.test.com/survery/A549/visit/2022-01-14/0 or more specifically https://id.test.com/survery/A549/visit/2022-01-14/sample/0
- The second observation of the third sample would then be https://id.test.com/survery/A549/visit/2022-01-14/sample/2/obs/1
- The first occurrence linked to the observation would be https://id.test.com/survery/A549/visit/2022-01-14/sample/2/obs/1/occ/0
- The fourth measurement linked to the occurrence would be https://id.test.com/survery/A549/visit/2022-01-14/sample/2/obs/1/occ/0/mof/3

If the event identifier is a simple string, then child events can be added by using a delimiter. Following the example above:

- A549-2022-01-14
- A549-2022-01-14-0
- A549-2022-01-14-2-1
- A549-2022-01-14-2-1:0
- A549-2022-01-14-2-1:0:3

Where we use a hyphen to indicate the continuation of a parent/child relationship and a colon to indicate a different type of entity.
Or you could just use UUIDs. We won't judge.

3.2 Occurrence

The occurrence[9] table contains all information related to occurrences. The how, when and where are contained in the associated event. The occurrence table contains the following minimum information:

- **eventID** The identifier of the associated event
- **occurrenceID** The occurrence unique identifier
- Enough information from the taxon[10] class to allow the what of the occurrence to be identified. This is usually either the scientificName or kingdom, phylum, … , genus, specificEpithet terms.

Any information from the occurrence[11], organism[12], geological context[13], identification[14] and taxon classes, as available and appropriate.

Generally, the location of the occurrence is the same as the location of the associated event. However, it is possible that that is not the case. For example, if an observer at lat/long -35.5103, 148.9457 observes a *Corvus mellori* at 35.511, 148.945 then the activity and the occurrence will have a different location and the occurrence table will need to contain the location of the occurrence. Empty location data in the occurrence table means that the location is inherited from the event. See 3.2.1 for a way of not doing this.

3.2.1 Preferred Practices

- The preferred location of location data is in the event table. If possible, create child events that represent the actual occurrence, rather than duplicate location data in the event and occurrence tables.
    - If both the event and the occurrence table both contain point location data, because the location of the event and the occurrence are clearly distinct (eg. sighting a bird 200m away from a hide location) then the event location is the position of the activity (an observation from the hide) and the occurrence location is the position of the organism (the bird). Be aware that this may lead to users missing either the event or the occurrence if they use narrow searches.
- It is theoretically possible for the eventDate of the event and occurrence to also be different[15]. The most likely cause of this is a coarse event with a number of occurrences attached to it. The preferred approach is to have observation child events attached to the original event, each with a separate eventDate.

3.3 Location

Generally, information from the location[16] class should be embedded in the event table, the location table or both.

The **locationID** is treated as a site identifier for survey-type information.

There are several forms of location information. Point location data is a typical form for a single observation and should contain at least the following terms:

[9] https://dwc.tdwg.org/terms/#occurrence
[10] https://dwc.tdwg.org/terms/#taxon
[11] https://dwc.tdwg.org/terms/#occurrence
[12] https://dwc.tdwg.org/terms/#organism
[13] https://dwc.tdwg.org/terms/#geologicalcontext
[14] https://dwc.tdwg.org/terms/#identification
[15] And not in the sense of [5].
[16] https://dwc.tdwg.org/terms/#location

- **decimalLatitude** The latitude of the point
- **decimalLongitude** The longitude of the point
- **geodeticDatum** The reference datum for the lat/long. If not specified, this will generally default to EPSG:4326 (WGS84)
- **coodinateUncertaintyInMeters** The uncertainty of the location
- **coodinatePrecision** The precision of the lat/long. Transforms between datums will generally respect the precision given.

Polygon location data can be used to identify the bounds of a site, a survey/expedition area, survey transect and the like. It should contain the following terms:

- **footprintWKT** The well-known-text [5] of the area.
- **footprintSRS** The spatial reference system for the footprint. If not specified, this will generally default to EPSG:4326 (WGS84)

As well as these terms, any other terms in the location class can be used.

3.3.1 Preferred Practices

- Sites should be given a consistent **locationID**.
- Include **country**, **stateProvince** and **locality** terms where possible, so that processing systems can sanity-check the coordinates provided and flag errors such as sign changes or zeroed coordinates. These terms also allow humans to quickly check the supplied data.
- The preferred form for datum and spatial reference system information is an EPSG code[17]. A datum/SRS should be explicitly provided.
- Polygon location data is often treated in mapping applications as the centroid point of the polygon. If **decimalLatitude**, **decimalLongitude**, **geodeticDatum** and **coordinateUncertaintyInMeters** are supplied along with the footprint then these can be used by mapping systems as a convenient dot and by mere humans as a location.

3.4 Extended Measurement or Fact

The extended measurement or fact extension[18] ties a measurement or assertion to either an event or an occurrence. The measurement or fact (MoF) table should contain as many of the following fields as are necessary:

- **eventID** The parent event identifier
- **occurrenceID** The parent occurrence identifier, if the MoF is associated with an occurrence, null if associated with an event.
- **xxxID** Any other identifiers that can be tied to MoFs, eg. the locationID if determining the location is part of some sort of process.
- **measurementID** A unique identifier for the measurement.
- **measurementType** A text description of the property being measured or the type of assertion being made. For example, "water acidity", "vegetation" or "wing span". These should not refer to the unit, if possible.
- **measurementTypeID**[19] A non-DwC term intended to describe the measurement type. Eg. http://qudt.org/vocab/quantitykind/IonicStrength
- **measurementValue** The result of the measurement or the face being asserted. Eg, using the measurement types above, "6.9", "myrtle-beech" or "10"

[17] https://epsg.io/
[18] https://rs.gbif.org/extension/obis/extended_measurement_or_fact.xml
[19] http://rs.iobis.org/obis/terms/measurementTypeID

- **measurementValueID**[20] A non-DwC term intended to describe the result of the measurement or assertion. Eg. http://anzsoil.org/def/au/asls/vegetation/myrtle-beech
- **measurementAccuracy** A statement, usually in the form of a plus/minus value. Eg. "0.01"
- **measurementUnit** The unit of measurement, Eg. "pH" or "cm". Assertions of fact have no unit.
- **measurementUnitID**[21] A non-DwC term indended to unambiguously describe the unit of measure. Eg. https://qudt.org/2.1/vocab/unit/PH
- **measurementDeterminedDate** The date (and time) the MoF was taken. If absent, the eventDate is used.
- **measurementDeterminedBy** The name of the person, group or organisation that determined the MoF
- **measurementRemarks** Free text of anything that you want.

3.4.1 Preferred Practices

- Use a well-known vocabulary for **measurementTypeID**
- Use the UCUM[22] standard for the text measurement units in **measurementUnit**.
- Use the QUDT[23] or NERC[24] vocabularies, if possible, for measurementUnitID
- Use the same rules for **eventDate** for **measurementDeterminedDate**. In general, a range will not be needed and instant date-times are expected.

4 Cases

Example files for each case can be found in https://archives.ala.org.au/archives/extendeddata

4.1.1 An Ecological Survey

https://archives.ala.org.au/archives/extendeddata/example1.zip
This survey follows the classical pattern of a number of sites which are consistently revisited and evaluated over a period of time. The survey is actually intended to collect data on drought levels and surface litter moisture content. However, it also generates occurrences
At each site, a number of sample plots have been surveyed and marked with pegs. Volunteers visit each corner and the centre of the plot, taking 1m x 1m ground litter samples at each location. The plants producing the litter are identified by leaf shape.
At the last (centre) location, the contents of the 1m x 1m square are sorted by species, bagged and returned to the lab. At the lab, the fresh weight and oven-dried weight of litter is measured, to give a measure of moisture content. A photo of the last sample, before it is bagged, is taken to ensure quality control.
At the start of each visit, the air temperature and relative humidity are measured. The Keetch-Byram Drought Index [9], a measure of how dry the underlying soil is, is also determined.
Any animals seen while walking the transect are also recorded. The protocol has been published on the institution's website. If possible, a photo is also taken.
The event core structure is as follows:

Event

[20] http://rs.iobis.org/obis/terms/measurementValueID
[21] http://rs.iobis.org/obis/terms/measurementUnitID
[22] https://ucum.org/
[23] https://qudt.org/
[24] http://vocab.nerc.ac.uk/

eventType	Parent eventType	eventDate	Point location	Polygon location	samplingProtocolID	occurrence
Survey		day range		Y		
SiteVisit	Survey	day		Y		
Sample	SiteVisit	hour range		Y	Y	
SubSample	Sample	instant	Y			Y
Observation	Sample	instant	Y			Y

Occurrence

Linked eventType	scientificName	Comments
SubSample	Y	Plants recorded in sampling frame
Observation	Y	Animals recorded while walking transect

Extended Measurement or Fact

Linked eventType	occurrenceID	measurementType	measurementUnit
Sample	N	KBDI	mm
SubSample	N	Soil Temperature	Cel
SubSample	N	Acidity	[pH]
SubSample	Y	Dry Weight	g

Multimedia

Linked eventType	occurrenceID
SubSample	**N**
Observation	Y

4.1.2 Identifiers

The identifiers for this dataset come from a RESTful service and, so, reflect the structure of the data. The identifiers can be decoded as

- http://id.test.com/survey/BVS - the survey itself, with the identifier BVS
- http://id.test.com/survey/BVS/site/SA/visit/0 - the first (index 0) visit to site A in the VBS survey.
- http://id.test.com/survey/BVS/site/SA/visit/0/sample/0 - the first full sample taken during the visit.
- http://id.test.com/survey/BVS/site/SA/visit/0/sample/0/2 - the third subsample taken in the sample
- http://id.test.com/survey/BVS/site/SA/visit/0/sample/1/observation/0 - the first observation of an animal made during the second sample of the first visit to site A in the BVS survey
- http://id.test.com/survey/BVS/site/SA/visit/0/occurrence/3 - the fourth occurrence seen during the first visit to site A
- http://id.test.com/survey/BVS/measurement/9 - the tenth measurement taken during the BVS survey.

In theory, visiting one of these URLs will provide data on activity, occurrence, measurement, etc. and there has been a script traversing the RESTful API, pulling the data out and turning it into a DwCA.

4.2 A Collecting Expedition

https://archives.ala.org.au/archives/extendeddata/example2.zip

A collecting expedition is where specimens are collected from a number of sites and returned to an institution for accession, identification and subsequent testing or treatment. Collecting expeditions can be less structured than the type of survey described in section 4.1.1 and can have multiple post-collection events that are conceptually the children of the collection event but which can happen a long time after the primary event.

The primary post-collection events are accession into a collection and condition reporting after accession. Accession and condition reporting events may also be associated with photographs of the specimen.

The event core structure is as follows:

Event							
eventType	**Parent eventType**	**eventDate**	**occurrence**	**Point location**	**Polygon location**	**eventName**	**eventDate outside parent range**
Expedition		day range			Y	Y	
SiteVisit	Expedition	day			Y	Y	
Collection	SiteVisit	day	Y	Y			
Identification	Collection	day					possible
Accession	Collection	day					Y
Preparation	Collection	day					Y
Condition	Collection	day					Y

Occurrence	
Linked eventType	**scientificName**
Collection	Y

Extended Measurement or Fact		
Linked eventType	**occurrenceID**	**measurementType**
Condition	Y	decay
Condition	Y	discoloration
Condition	Y	pests

Multimedia	
Linked eventType	**occurrenceID**
Accession	Y
Condition	Y

5 Glossary

DwCA: Darwin Core Archive. A package of related biodiversity data.
OBIS: Ocean Biodiversity Information System

6 Bibliography

[1] P. Leach, M. Mealling and R. Salz, "RFC 4122: A Universally Unique IDentifier (UUID) URN Namespace," July 2005. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc4122.

[2] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool, 2011.

[3] L. Sauermann, R. Cyganiak, D. Ayers and M. Völkel, "Cool URIs for the Semantic Web," 2008. [Online]. Available: https://www.w3.org/TR/cooluris/.

[4] GBIF, "Darwin Core Archives – How-to Guide," [Online]. Available: https://ipt.gbif.org/manual/en/ipt/2.5/dwca-guide.

[5] M. B. Jones, M. O'Brien, B. Mecum, C. Boettiger, M. Schildhauer, M. Maier, T. Whiteaker, S. Earl and S. Chong, "Ecological Metadata Language version 2.2.0.," 2019. [Online]. Available: https://eml.ecoinformatics.org/.

[6] Unicode Inc., "UTF-8, UTF-16, UTF-32 & BOM," 2022. [Online]. Available: https://www.unicode.org/faq/utf_bom.html.

[7] Ocean Biodiversity Information System, "The OBIS Manual," [Online]. Available: https://manual.obis.org/.

[8] International Standards Organisation, "ISO/IEC 13249-3:2016 Information technology — Database languages — SQL multimedia and application packages — Part 3: Spatial," 2016. [Online]. Available: https://www.iso.org/standard/60343.html.

[9] J. J. Keetch and G. M. Byram, "A Drought Index for Forest Fire Control," 1968.

[10] A. Einstein, "On the Electrodynamics of Moving Bodies," *Annalen Phys.,* vol. 17, pp. 891-921, 30 June 1905.