How to create a Darwin Core Archive for scientists

Luke Marsden (lukem@met.no)

January 18, 2024

Contents

1 What is Darwin Core

Darwin Core is a data standard originally developed for biodiversity informatics, though this has expanded to be useful for any type of data where you have data associated with one or a list of organisms. Darwin Core includes

- Darwin Core terms: A controlled vocabulary of terms - https://dwc.tdwg.org/terms/
- Darwin Core Archive: A FAIR-compliant data format

2 What is a Darwin Core Archive

A Darwin Core Archive (DwCA) is a self-describing dataset for taxonomic (species) data, sampling event data and associated traits or measurements. It consists of one or more data tables (CSV files) and 2 XML files, one (meta.xml) that describes how the files are organised and a second (eml.xml) that provides the metadata describing the dataset as a whole. They are zipped together to create the Darwin Core Archive (DwCA) (Figure 1). Don't worry if you don't know what an XML file is. You don't need to create this yourself. This is done for you using the Integrated Publishing toolkit as described in section 7.2.

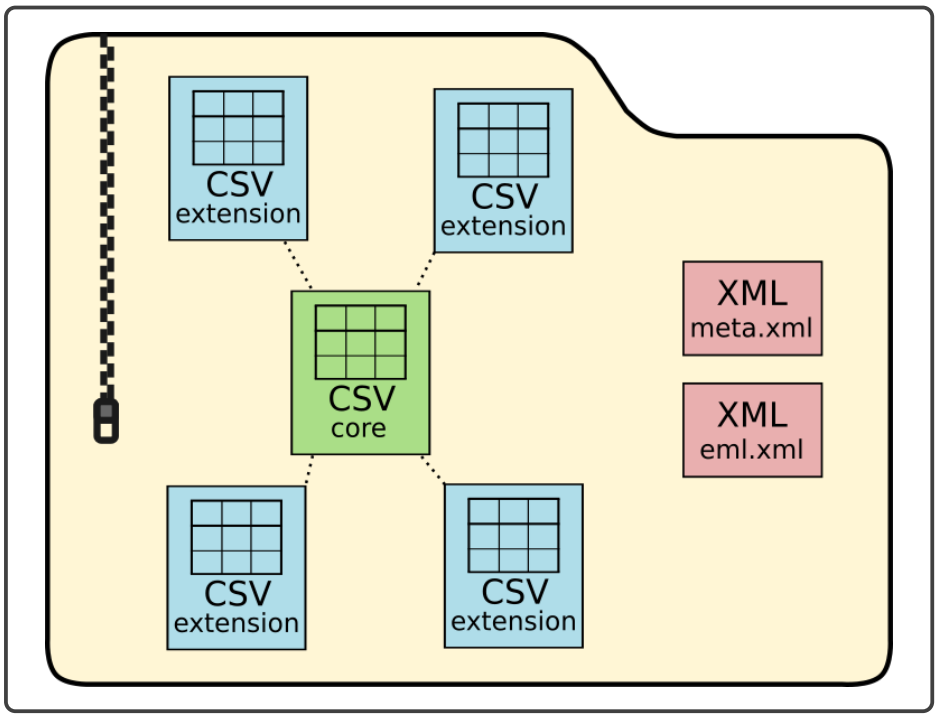

Figure 1: Visualisation of a Darwin Core Archive, portrayed using the star schema. The central event core can be surrounded by zero or many extension tables. It also contains a meta.xml file that describes what columns each CSV contains and links them to the term in a controlled volcabulary, and an eml.xml file that provides metadata that describes the dataset as a whole. They are zipped together to create a Darwin Core Archive.

Each CSV file contains a number of rows, and every row has its own unique ID (Figure 2). These IDs are used to link the CSV files together. Every row in every *extension* CSV file must include the ID of one row in the *core* CSV file. This is called the *star schema* and you can think of the core as the centre of the star. Note that a DwCA can only include a single level of extensions. In other words, one cannot include an extension to an extension file.

For more information on DwCA, see https://ipt.gbif.org/manual/en/ipt/2.5/dwca-guide.

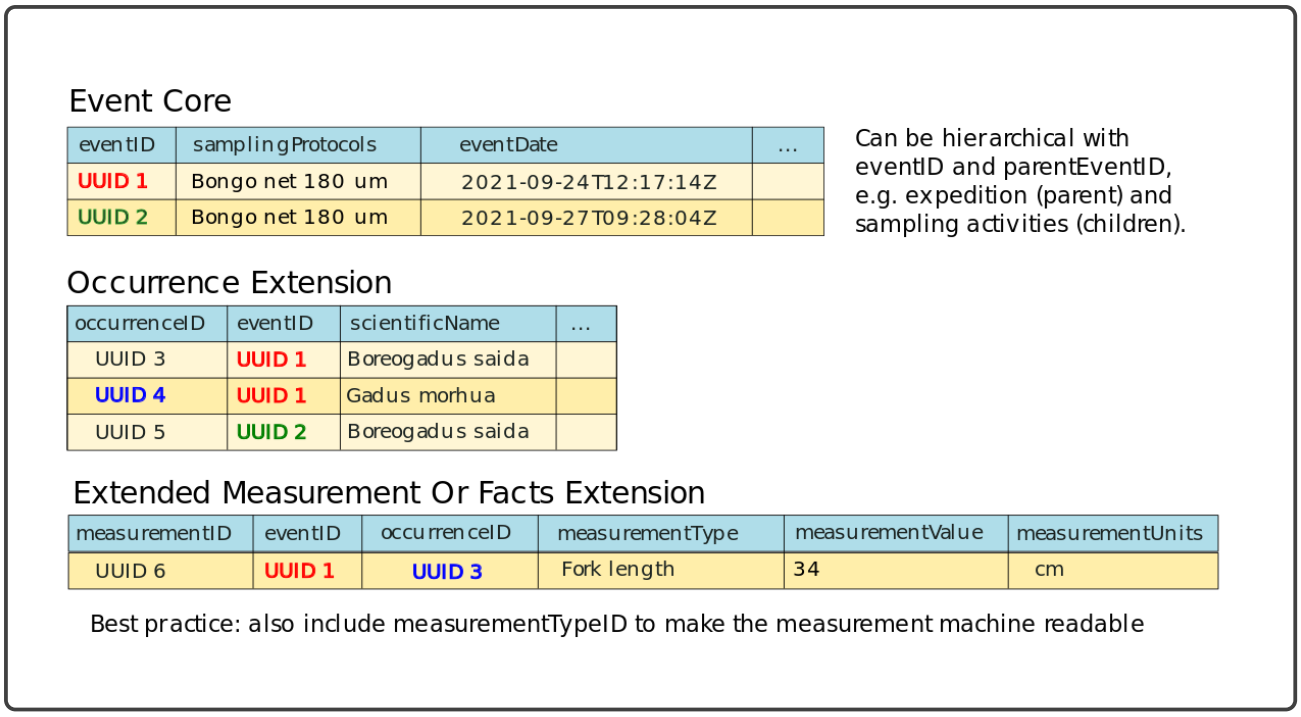

Figure 2: Example of a Darwin Core Archive with an Event Core and 2 extensions; an Occurrence Extension and a Extended Measurement Or Facts Extension. Note that this example does not include all the required columns for each core or extension.

Visual learners might want to watch this video about DwCAs: https://www.youtube.com/watch?v=1Fuq8VZW_4c

2.1 Why do we use multiple CSV files?

This is a common question, but we are not trying to make things needlessly complicated! This allows many-to-one relationships to be logged, for example multiple occurrences (observations of an organism or group of the same organism) logged for a single sampling event.

This method is also more efficient. The metadata related to the sampling event (coordinates, date, time etc) only needs to be logged once per sampling event instead of for each occurrence. This can save a lot of time and space, particularly if you have a lot of occurrences for each sampling event!

3 Darwin Core terms and other controlled vocabularies

Darwin Core includes a controlled vocabulary of terms for sharing information about biological diversity and associated data. Most (but not all) of the column headers you will include in your CSV files in your DwCA will be Darwin Core terms. Other terms will be from other controlled vocabularies.

A full list of Darwin Core terms can be found at https://dwc.tdwg.org/terms/.

4 GBIF, OBIS and a network of data

There are a number of data centres and services that manage Darwin Core data. GBIF (Global Biodiversity Information Facility) is the largest. Most (all?) of the other Darwin Core services

make their data available via GBIF as well as their own platform. So if you publish your data with any of the below services, your data will also be available via GBIF.

The table below is just a few of the relevant platforms that contribute to this network of data.

Table 1: A network of Darwin Core data. This table is just a small selection of the services available.

Name	Website	Comment
OBIS	https://obis.org/	Aims to include all marine data. The OBIS community have been pushing the DwC standards forwards to build better functionality for scientific data.
iNaturalist	https://www.inaturalist.org/	For citizen science, nature enthusiasts and researchers. Offer some great apps like Seek that you can use on your mobile phone for taking pictures, identifying the organism and publishing the data https://www.inaturalist.org/pages/seek_app
Living Norway	https://livingnorway.no/	Norwegian ecological data project
Artsdatabanken	https://www.artsdatabanken.no/	Service for collecting, organizing, and disseminating data related to Norwegian flora and fauna

5 Cores and extensions

Which cores and extensions should you include in your DwCA?

For most scientific data, you should be include and 'Event Core', where 1 row is 1 sampling event. This can also be hierarchical, using eventID and parentEventID (Figure 3). You can include as many levels in the hierarchy as you wish. Every extension should then also include an eventID column where you input the ID of the sampling event that the data are related to. Refer to the child if using parent-child relationships in your event core.

https://rs.gbif.org/core/dwc_event_2022-02-02.xml

You will most likely also have an occurrence extension, where one row is an observation of an organism of group of organisms of the same species.

https://rs.gbif.org/core/dwc_occurrence_2022-02-02.xml

Measurements related to the sampling event or occurrence (e.g. mass of your organism) go in an Extended Measurement or Facts extension.

https://rs.gbif.org/extension/obis/extended_measurement_or_fact_2023-08-28.xml

If you have material samples (soil, faeces) that you have taken your can log them using a Material Sample extension.

https://rs.gbif.org/extension/ggbn/materialsample.xml

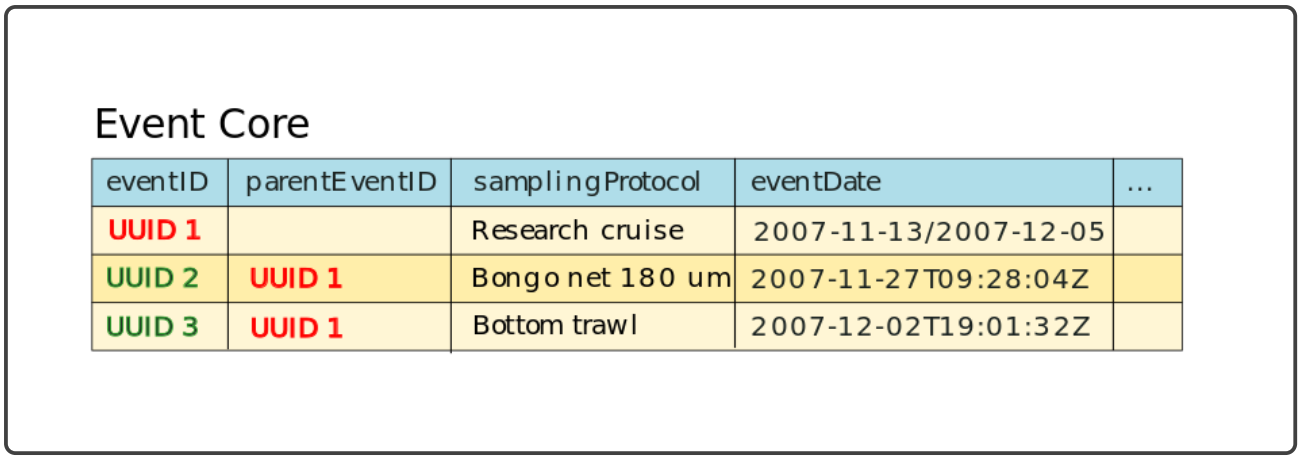

Figure 3: Example of a hierarchical event core. The research cruise is the parent of the children sampling activities. This is shown using the eventID and parentEventID columns.

If you have associated media you can refer to them using a Simple Multimedia Extension, and publish the media elsewhere. It is possible to publish these with GBIF too.

http://rs.gbif.org/extension/gbif/1.0/multimedia.xml

If you need to relate a row in one extension to a row in another extension, you can use the Resource Relationship Extension.

https://rs.gbif.org/extension/dwc/resource_relationship_2022-02-02.xml

And more extensions that are registered with GBIF can be found here.

https://rs.gbif.org/extensions.html

6 Which columns are required in each extension?

It can be difficult to know which columns to include and what the minimum requirements are. Unfortunately, this information is not easy to find online. You can contact the data centre that you are publishing your data with.

Every core and extension should include an ID column that is specific to that extension or core (eventID for Event Core, occurrenceID for Occurrence Extension, measurementID for Extended Measurement or Facts Extension, etc). Every extension should also include the ID column for the core, e.g. if you have an Event Core, every extension must include an eventID column. Best practice is to use universally unique identifiers (UUIDs) for the IDs. You can generate them in most programming languages or using tools online such as https://www.uuidgenerator.net/version4.

I have also developed the Nansen Legacy spreadsheet template generator for Darwin Core Archives that has required and recommended columns for different cores and extensions included. Even if you would rather not use spreadsheets, you could use this to see which columns you should include.

https://www.nordatanet.no/aen/template-generator/config%3DDarwin%20Core

If you notice any issues with this or want to see something else added, please get in touch,

or better still, raise an issue on GitHub.

https://github.com/SIOS-Svalbard/Nansen_Legacy_template_generator

7 How to create a Darwin Core Archive

To create a Darwin Core Archive you first have to create the CSV files. Then, you can use the Integrated Publishing Toolkit (IPT), developed by GBIF, to turn them into a Darwin Core Archive.

7.1 Creating the CSV files

You could use the Nansen Legacy template generator to help you with this. This will help you create a spreadsheet with a separate sheet for each core and extension (Figure 4). The descriptions for each term will appear as notes each time you select a cell to help you fill it in. I advise that you export each sheet to separate CSV files before you proceed with creating the Darwin Core Archive.

This video shows you how to use to template generator and how to publish the data afterwards: https://www.youtube.com/watch?v=DbvlwnYXuPU

Alternatively, you can create CSV files in any of way that you prefer. You can go to https://rs.gbif.org/extensions.html to see which cores and extensions you can use and which terms you can include in each. Note that it does not tell you which terms are required. This varies between data centres - GBIF has slightly different minimum requirements to OBIS for example for certain extensions.

7.2 Creating a DwCA from your CSVs

Once you have your CSV files, creating a DwCA is easy. You can using the Integrated Publishing Toolkit (IPT), developed by GBIF, to create the DwCA and also publish it.

Here is a map of places that have an IPT installed https://www.gbif.org/ipt. Choose one, contact them and ask for login details to their node.

For most people in Norway, good choices are the Norwegian Marine Data Centre (operated by IMR) - datahjelp@imr.no, or GBIF Norway - helpdesk@gbif.no.

For how to use the IPT, watch the last 3 minutes of this video: https://www.youtube.com/watch?v=DbvlwnYXuPU&t=280s. The XML files are created through the IPT for you - you just need to upload your CSV files, select (map them to) which core or extension they are, and include metadata using a form.

This is a more extensive video tutorial on how to use the IPT created by GBIF (24 minutes): https://www.youtube.com/watch?v=eDH9IoTrMVE

8 Making your data available via SIOS

Data relevant to Svalbard should be available via the SIOS data access portal https://sios-svalbard.org/metsis/search?f%5B0%5D=dataset_level%3ALevel-1

If you have published your data with the Norwegian Marine Data Centre, they contribute to SIOS so the data will be made available via the SIOS.

If you publish your data elsewhere (e.g. with GBIF Norway) you need to link your data to SIOS manually. You can use the metadata collection form on SIOS to do this. You will need to log in first. https://sios-svalbard.org/metadata-collection-form

9 Citing your data in your paper

Cite your data just as you would cite any other scientific publication - in your list of references. You can also mention the data in a data availablity statement if your chosen journal requires one, but this should be as well as (not instead of) including the data in your list of references.

The recommended citation can be seen on the landing page of the dataset in the data centre you chose to publish with (GBIF or NMDC most likely).

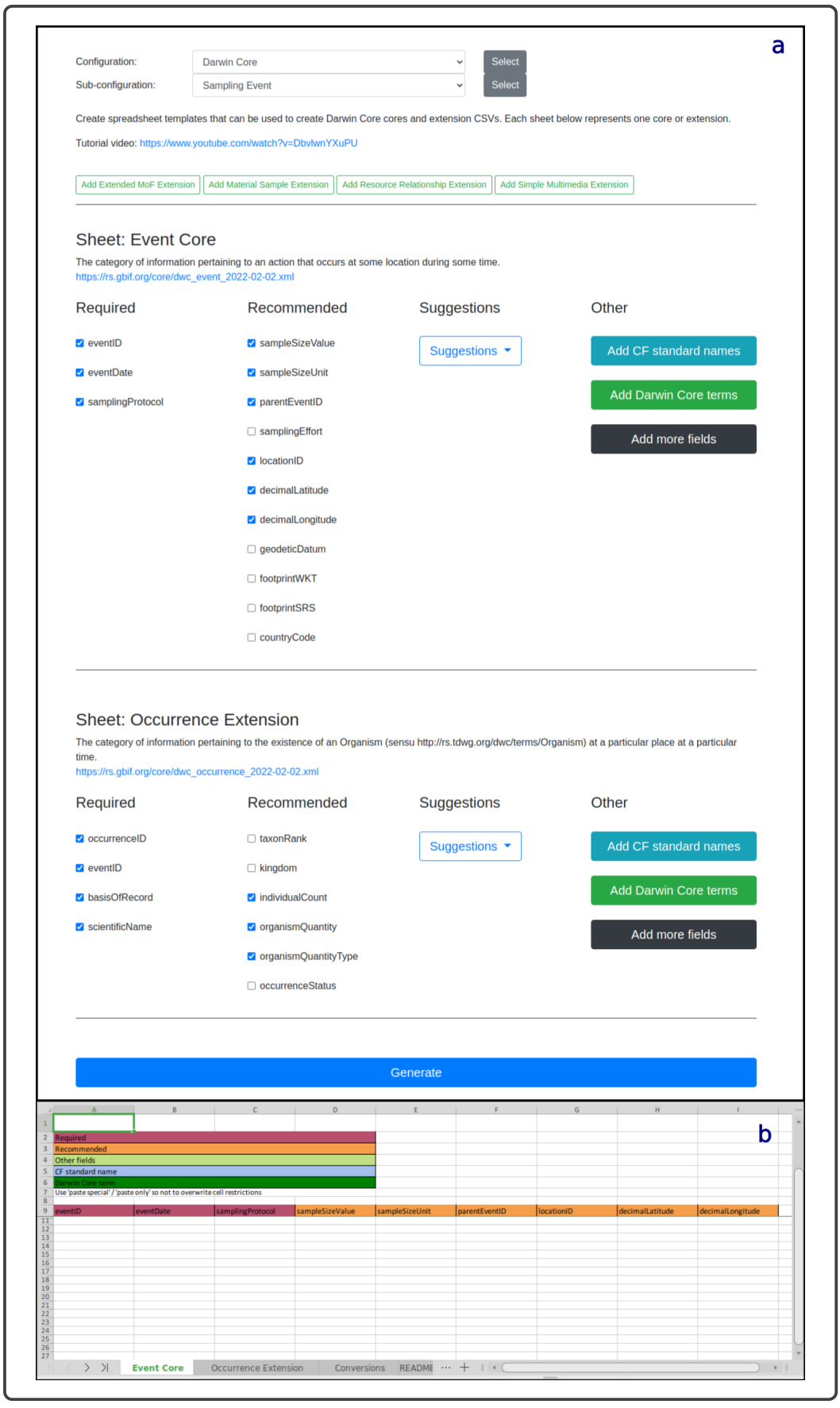

Figure 4: The Nansen Legacy template generator (a) and an example template (b)