# A Discussion of Darwin Core with respect to The Paleobiology Database

Version 1.0: 2016-03-02

*Michael McClennen  
Informaticist and Senior Software Developer*

The Darwin Core standard is of obvious interest to us, since it is by far the most widely used standard for data interchange regarding occurrences and specimens of organisms. Unfortunately, it is a very limited standard, with many gaps and ambiguities which will be discussed below. Nevertheless, it is the “only game in town” for data interchange with a wide variety of other databases such as iDigBio and GBIF.

## Current capability of The Paleobiology Database

At the present time (March, 2016) the PaleoBioDB data service does not provide output in the Darwin Core vocabulary. The data service codebase does, however, have the built-in capability of providing output in multiple vocabularies. Output using Darwin Core fields in the comma-separated value format (CSV) could be enabled in just a day or two, *provided* that field and value mappings are first decided upon and any ambiguities are resolved. As discussed below, this may take quite a bit of effort.

Output files in CSV format could then be manually assembled together with a suitable metadata file into Darwin Core Archive files according to the Darwin Core standard[[1]](#endnote-1). Direct output of Darwin Core Archive files in response to data service requests would require new code to be written, which would involve about 1-2 weeks of additional programming effort.

Data input to The Paleobiology Database is currently limited to manual entry using the input forms provided by the Classic interface. Work is underway to change this, by extending the data service to provide for upload of data in both single-record and bulk form. This will probably take at least a year to come to fruition. This capability, when it arrives, will allow for input translation from multiple vocabularies. Similarly to the data output side of things, it will be relatively straightforward to enable input of Darwin Core data *provided* that field and value mappings are first decided upon and any ambiguities resolved.

The following sections discuss in greater detail the issues that must be dealt with. The tables of field names included below indicate areas of ambiguity, as well as PaleoBioDB fields that are not representable in the current version of Darwin Core.

## Limitations of the Darwin Core standard

The limitations of Darwin Core with respect to The Paleobiology Database fall into several broad categories:

1. Ambiguity of interpretation
2. Missing field names
3. Missing data classes
4. Ambiguity in the relationship between fields and classes

Each of these limitations implies certain decisions that must be made before we can successfully transfer data to our partner organizations using Darwin Core or receive data from them.

### Ambiguity of interpretation

Perhaps the most important limitation of the Darwin Core standard is that it does not enforce any rules on the interpretation of field values. This can be considered both a strength, in that it provides broad flexiblity in the kinds of data that can be represented, and a weakness, in that there is no guarantee that the recipient of data will understand the field values as they are received or that it will interpret them in the manner intended by the sender of the data.

As a result, simply parceling our data into the appropriate Darwin Core fields is not sufficient to enable a useful interchange of information. For each database or data aggregator to which we wish to send information, we will have to work closely with somebody in that organization to find out exactly how their database will interpret each Darwin Core field and decide how we should encode or translate our data values in order that they will be properly understood on the other end. This translation will have to be customized for each separate recipient of our data.

Similarly, if we wish to import data using Darwin Core we will need to work closely with the people who will be exporting that data to find out the ranges of values that each field will take on, so that we can properly translate them into values that match the schema of our local database. Again, this translation will have to be customized for each separate source from which we wish to import data.

This issue is not really a problem with Darwin Core, it is the unavoidable outcome of the differences between the data schema that we use and the data schemas used by our various partner organizations.

### Missing field names

Because Darwin Core was defined by people who were primarily concerned with modern biodiversity, it is focused on extant organisms and on specimens and samples from the present time. Consequently, there are entire categories of information that are simply not representable in Darwin Core as it now exists, and others that will need to be crammed into single fields with no guide as to how they should be represented or interpreted. The categories that will cause the most trouble for us include:

* Geochronology
* Paleolocation
* Paleoenvironment
* Lithology
* Detailed stratigraphy

Of these, I think the first is the most crucial for us. A Darwin Core record can (clumsily) include a chronostratigraphic interval, but it cannot include a direct date or range of dates from deep time. This information is a vital component of our data, and we will need to work with the institutions with which we are exchanging data in order to figure out some way to handle it.

The other categories we may be able to ignore, since most of the databases to which we are sending data will have no way of representing these data values to their users. My sense is that our partner institutions are never going to be able to represent every dimension of our data, nor are we going to be able to represent every dimension of theirs. Rather, we should focus on sharing some subset of information that fits into the intersection of the data schemas in question, and make clear to end users that for more detailed information about the records of interest they must directly query each record’s source database.

If we choose not to follow this course of action, then we will have to figure out ways of working around these missing fields. Fortunately, there are some “catch-all” fields in the Darwin Core standard, which can potentially be used to encode field values that are not otherwise representable. Examples include *dynamicProperties* and the various *remarks* fields. However, this will require coordination with our partner institutions so that they know the particulars of how we are encoding these values.

## Missing data classes

In my opinion, the lack of a container class for other records is the single biggest problem with the Darwin Core standard. The set of classes defined by the current version of the standard is as follows (with subclasses)[[2]](#endnote-2):

* Occurrence
* Organism
* MaterialSample (LivingSpecimen, PreservedSpecimen, FossilSpecimen)
* Event (HumanObservation, MachineObservation)
* Location
* GeologicalContext
* Identification
* Taxon
* MeasurementOrFact

As you can see, this set of classes is able to describe almost all of our data records *except for collections*. There is no provision for representing (for example) a group of occurrence records as a separate object with its own attributes, which in our case would include location, stratigraphy, collection methods, and so on.

There are two fields called *datasetID* and *datasetName* which could be used in occurrence records to indicate which collection each occurrence belongs to, but there is no way to indicate the values of attributes that are common to the collection other than by duplicating these values in each individual occurrence record. This is not an insurmountable problem, but it does complicate our use of Darwin Core and multiplies the size of the data files we will be generating and incorporating.

This hole in the Darwin Core standard is quite puzzling to me. I strongly suspect that it is also a problem for other users of the standard, who would surely like to be able to represent groups of samples, observations, events, etc. as separate instances of a container class where each such record represents the attributes common to a group. To whatever extent we are able to provide input to TDWG, I strongly suggest that we advocate the addition of a basic container class to the standard. The easiest way to do this would be to add *Dataset* to the list of standard classes, since *datasetID* and *datasetName* are already there as fields. There should also be a *parentDatasetID* field, to allow for groups of groups, and *datasetType* to allow for multiple kinds of groups.

* + 1. **Other missing classes**

There are two other missing classes that are relevant to our own data in particular. We could potentially advocate for their addition to the standard as well, but I regard them as of lower priority than the container class.

**Taxonomic opinions**

Darwin Core lacks a representation for taxonomic opinions. This is probably a moot point for the present, since to my knowledge we are the only database in which taxonomic opinions play a serious role. If we ever wish to exchange data with some other database that also deals with taxonomic opinions, we will have to negotiate some side channel by which to communicate the opinion data, or else some convoluted way of encoding it by using special values in the existing Darwin Core fields.

**Bibliographic references**

There are at least four different ways to indicate bibliographic references in Darwin Core: *associatedReferences*, *identificationReferences*, *nameAccordingToID*,and *dcterms:bibliographicCitation*. These can be used to communicate either of the following:

* A list of reference identifiers
* A bibliographic citation as free text

There is no provision in Darwin Core for communication of a bibliographic citation as a structured record with fields for author, title, publication, etc. We can probably work around this by transmitting unique identifiers such as DOIs. Again, this would have to be negotiated with each partner organization separately.

As a side note, the multiplicity of these fields indicates yet another sloppiness in the definition of Darwin Core. It looks as though the issue of citation was overlooked at first, and clumsily added in piecemeal fashion later.

## Ambiguity in the relationship between fields and classes

The Darwin Core documentation associates each field with one particular class, but it is not clearly specified whether each field is limited to use with its particular class. As far as I can tell from looking over the historical notes, the relationship between fields and classes was formerly strictly enforced. This association seems to have been loosened in recent years, so that any record is now allowed to contain any fields regardless of class.

This is important for us, since it is difficult to express our data properly if these relationships are strictly enforced. Older software systems may in fact enforce these restrictions. If this is the case for some or all of the systems to which we plan to send data, we will have to figure out some workarounds. This would necessitate extra programming work on our end, and I cannot currently estimate how much extra work would be required.

## PaleoBioDB output and its relationship to Darwin Core

The following subsections describe the fields that are output by the PaleoBioDB data service, and discuss some of the issues that may come up in translating them into Darwin Core. It is organized by data class, and within each class according to the output blocks produced by the data service.

Whenever *n/e* appears, it indicates a PaleoBioDB field that has no direct equivalent in Darwin Core. As discussed above, we will either have to leave this data out or negotiate some arrangement with the data recipient whereby we encode certain data values using the “catch-all” fields such as *dyna*m*icProperties* and the various *remarks* fields.

Wherever more than one term appears in the right-hand column, there is more than one Darwin Core term that could correspond to the PaleoBioDB field name. We will again have to negotiate with the data recipient as to which field should be used.

## Occurrences

We start with the occurrence record since it is the core data type for the PaleoBioDB. The following table describes the *basic* block produced by the data service when listing occurrences.

|  |  |  |
| --- | --- | --- |
| Block | PaleoBioDB field name | Darwin Core alternatives |
| *basic* | occurrence\_no | occurrenceID |
| *basic* | record\_type | dcterms:type |
| *basic* | collection\_no | datasetID |
| *basic* | reid\_no | identificationID |
| *basic* | identified\_name | scientificName |
| *basic* | identified\_rank | taxonRank |
| *basic* | identified\_no | scientificNameID |
| *basic* | accepted\_name | scientificName  acceptedNameUsage  associatedTaxa |
| *basic* | accepted\_rank | taxonRank |
| *basic* | accepted\_no | scientificNameID  acceptedNameUsageID |
| *basic* | accepted\_attr | scientificNameAuthorship |
| *basic* | early\_interval | earliestAgeOrLowestStage  earliestEonOrLowestEonothem  earliestEpochOrLowestSeries  earliestEraOrLowestErathem  earliestPeriodOrLowestSystem |
| *basic* | late\_interval | latestAgeOrHighestStage  latestEonOrHighestEonothem  latestEpochOrHighestSeries  latestEraOrHighestErathem  latestPeriodOrHighestSystem |
| *basic* | early\_age | n/e |
| *basic* | late\_age | n/e |
| *basic* | ref\_author | recordedBy |
| *basic* | ref\_pubyr | n/e |
| *basic* | reference\_no | associatedReferences |

The first four fields are relatively straightforward, except for the issue noted above about the lack of a container class. We can use *datasetID* to indicate collection membership, but there is no way to separately communicate the attributes of a collection as an individual record.

Darwin Core can easily represent the taxonomic name associated with an occurrence, along with a unique identifier associated with the name. It can also represent reidentifications, using the *identificationID* field. The main complication is that the PBDB occurrence includes both an “identified name”, which is the name specified in the reference from which the occurrence was entered, and an “accepted name” which is dynamically derived from the taxonomic hierarchy. Subsequent changes in the taxonomic or nomenclatural status of a name or changes in spelling will cause these two names to differ.

It is possible to represent both of these names in Darwin Core, using the *scientificName* field to represent the “identified name” and *acceptedNameUsage* to represent the “accepted name”. However, the databases to which we will be exporting records most likely have the capability of representing only a single taxonomic name for each occurrence. In that case, we will have to decide whether to send the “identified name” or the “accepted name” and leave out the other one.

The chronostratigraphic information represents another tricky problem. The GeologicalContext fields of Darwin Core can represent interval names, but there ate not generic fields for “earliest interval” and “latest interval”. On top of that, Darwin Core has no way at all to represent the “early\_age” and “late\_age” fields. Once again, exactly how we represent this information will have to be negotiated with our partner organizations.

The following table represents the other output blocks that are available for occurrences. Some of the output blocks that can be selected when listing occurrences properly belong to the collection type, and are listed in the next section.

|  |  |  |
| --- | --- | --- |
| Block | PaleoBioDB field name | Darwin Core alternatives |
| *ident* | genus\_name | genus |
| *ident* | genus\_reso | n/e |
| *ident* | subgenus\_name | subgenus |
| *ident* | subgenus\_reso | n/e |
| *ident* | species\_name | specificEpithet |
| *ident* | species\_reso | n/e |
|  |  |  |
| *class* | phylum | phylum |
| *class* | class | class |
| *class* | order | order |
| *class* | family | family |
| *class* | genus | genus |
| *class* | subgenus | subgenus |
|  |  |  |
| *plant* | plant\_organ | n/e |
|  |  |  |
| *coords* | lng | decimalLongitude |
| *coords* | lat | decimalLatitude |
|  |  |  |
| *abund* | abund\_value | organismQuantity |
| *abund* | abund\_unit | organismQuantityType |

The main complication here is determining how to represent taxonomic abbreviations such as *aff.* *cf.* *n. gen.* and so forth. There is no specified way to do this in the Darwin Core standard, so it will again have to be negotiated individually with the various partner organizations.

## Collections

Because of the aforementioned lack of a collection class, we will have to represent the attributes of collections as attributes of the individual occurrences. The only alternative would be to negotiate with one or more of our partner organizations an informal extension of the standard to include a *Dataset* class. Be that as it may, the following table describes the basic fields of a PaleoBioDB collection record along with potential translations into Darwin Core.

|  |  |  |
| --- | --- | --- |
| Block | PaleoBioDB field name | Darwin Core alternatives |
| *basic* | collection\_no | datasetID |
| *basic* | record\_type | dcterms:type |
| *basic* | formation | formation |
| *basic* | lng | decimalLongitude |
| *basic* | lat | decimalLatitude |
| *basic* | collection\_name | datasetName |
| *basic* | collection\_subset | n/e |
| *basic* | n\_occs | n/e |
| *basic* | early\_interval | earliestAgeOrLowestStage  earliestEonOrLowestEonothem  earliestEpochOrLowestSeries  earliestEraOrLowestErathem  earliestPeriodOrLowestSystem |
| *basic* | late\_interval | latestAgeOrHighestStage  latestEonOrHighestEonothem  latestEpochOrHighestSeries  latestEraOrHighestErathem  latestPeriodOrHighestSystem |
| *basic* | early\_age | n/e |
| *basic* | late\_age | n/e |
| *basic* | ref\_author | recordedBy |
| *basic* | ref\_pubyr | n/e |
| *basic* | reference\_no | associatedReferences |

There are no issues arising here that have not been dealt with in the previous section. The following table represents the other output blocks available when listing collections through the data service.

|  |  |  |
| --- | --- | --- |
| Block | PaleoBioDB field name | Darwin Core alternatives |
| *loc* | cc | countryCode |
| *loc* | state | stateProvince |
| *loc* | county | county |
| *loc* | latlng\_basis | locationAccordingTo  locality |
| *loc* | latlng\_precision | coordinatePrecision  locality |
| *loc* | geogscale | locationRemarks |
| *loc* | geogcomments | locationRemarks |
|  |  |  |
| *paleoloc* | paleolng | n/e |
| *paleoloc* | paleolat | n/e |
| *paleoloc* | geoplate | n/e |
| *paleoloc* | paleomodel | n/e |
|  |  |  |
| *prot* | protected | locationRemarks |
|  |  |  |
| *strat* | formation | formation |
| *strat* | geological\_group | group |
| *strat* | member | member |
| *stratext* | stratscale | lithoStratigraphicTerms |
| *stratext* | zone | lowestBioStratigraphicZone  highestBioStratigraphicZone |
| *stratext* | localsection | lithoStratigraphicTerms |
| *stratext* | localbed | bed |
| *stratext* | localorder | lithoStratigraphicTerms |
| *stratext* | regionalsection | lithoStratigraphicTerms |
| *stratext* | regionalbed | bed |
| *stratext* | regionalorder | lithoStratigraphicTerms |
| *stratext* | stratcomments | lithoStratigraphicTerms |
|  |  |  |
| *lith* | lithdescript | lithoStratigraphicTerms |
| *lith* | lithology1 | lithoStratigraphicTerms |
| *lithext* | lithadj1 | lithoStratigraphicTerms |
| *lith* | lithification | lithoStratigraphicTerms |
| *lith* | minor\_lithology | lithoStratigraphicTerms |
| *lithext* | fossilsfrom | lithoStratigraphicTerms |
| *lith* | lithology2 | lithoStratigraphicTerms |
| *lithext* | lithadj2 | lithoStratigraphicTerms |
| *lith* | lithification2 | lithoStratigraphicTerms |
| *lith* | minor\_lithology2 | lithoStratigraphicTerms |
| *lithext* | fossilsfrom2 | lithoStratigraphicTerms |
|  |  |  |
| *env* | environment | n/e |
| *env* | tectonic\_setting | n/e |
| *env* | geology\_comments | n/e |
|  |  |  |
| *methods* | collection\_type | samplingProtocol |
| *methods* | collection\_methods | preparations |
| *methods* | museum | institutionCode |
| *methods* | collection\_coverage | samplingProtocol |
| *methods* | rock\_censused | samplingProtocol |
| *methods* | collectors | recordedBy |
| *methods* | collection\_dates | eventDate |
| *methods* | collection\_comments | eventRemarks |
|  |  |  |
| *rem* | collection\_aka | n/e |

The obvious problems here derive from the complete absence of fields for representing paleolocation and paleoenvironment, and a single field for representing all lithostratigraphic information. Once again, we will either have to leave all or most of this data out, or negotiate with our partner organzations a method of encoding it that they can unpack and interpret.

## Taxa

The following table describes the *basic* block produced by the data service when listing taxa.

|  |  |  |
| --- | --- | --- |
| Block | PaleoBioDB field name | Darwin Core alternatives |
| *basic* | taxon\_no | taxonID |
| *basic* | orig\_no | taxonConceptID |
| *basic* | record\_type | dcterms:type |
| *basic* | taxon\_rank | taxonRank |
| *basic* | taxon\_name | scientificName |
| *basic* | taxon\_attr | scientificNameAuthorship |
| *basic* | common\_name | vernacularName |
| *basic* | difference | n/e |
| *basic* | accepted\_no | acceptedNameUsageID |
| *basic* | accepted\_rank | n/e |
| *basic* | accepted\_name | acceptedNameUsage |
| *basic* | senpar\_no | parentNameUsageID |
| *basic* | immpar\_no | parentNameUsageID |
| *basic* | reference\_no | namePublishedInID |
| *basic* | is\_extant | n/e |
| *basic* | n\_occs | n/e |

The Darwin Core standard is fairly complete here, with the glaring exception that there is no way to differentiate extant taxa from those that are extinct. Once again, we will have to negotiate some special way to represent this.

One complication that may have to be dealt with is that a PaleoBioDB taxon record specifies both an immediate parent taxon and a senior parent taxon. The latter is defined as the senior synonym of the immediate parent if the immediate parent is a junior synonym. Darwin Core only provides the ability to represent a single parent taxon. I suspect that most of the other databases to which we will be sending information will want the “senior parent”, and so that is the attribute that should be translated to *parentNameUsageID.*

The following table lists the other output blocks that are available when listing taxa through the PaleoBioDB data service.

|  |  |  |
| --- | --- | --- |
| Block | PaleoBioDB field name | Darwin Core alternatives |
| *app* | firstapp\_max\_ma | n/e |
| *app* | firstapp\_min\_ma | n/e |
| *app* | lastapp\_max\_ma | n/e |
| *app* | lastapp\_min\_ma | n/e |
|  |  |  |
| *class* | kingdom | kingdom |
| *class* | phylum | phylum |
| *class* | class | class |
| *class* | order | order |
| *class* | family | family |
| *class* | genus | genus |
| *class* | subgenus | subgenus |
|  |  |  |
| *img* | image\_no | associatedMedia |
|  |  |  |
| *ecospace* | life\_habit | n/e |
| *ecospace* | diet | n/e |
|  |  |  |
| *taphonomy* | composition | n/e |
| *taphonomy* | architecture | n/e |
| *taphonomy* | thickness | n/e |
| *taphonomy* | reinforcement | n/e |
|  |  |  |
| *pres* | preservation | n/e |

There is no way to represent most of this information in Darwin Core, with the exception of the taxonomic classification.

## Opinions

As discussed above, taxonomic opinions cannot be represented in Darwin Core as it currently exists.

## References

As discussed above, bibliographic references can only be represented in Darwin Core by unique identifiers or as free-text citations. There is no way to split out the individual attributes such as authors, title, publication, page numbers, and so forth. This represents a complication that will have to be negotiated with our partner organizations.

|  |  |  |
| --- | --- | --- |
| Block | PBDB field name | Darwin Core alternatives |
| *basic* | reference\_no | resourceID |
| *basic* | record\_type | dcterms:type |
| *basic* | reftitle, etc. | dcterms:bibliographicCitation |

## PaleoBioDB input and its relationship to Darwin Core

The issue of input will have to wait until our new input system has been further developed.

## Endnotes

1. <http://rs.tdwg.org/dwc/terms/guides/text/index.htm> (Darwin Core Text Guide) [↑](#endnote-ref-1)
2. <http://rs.tdwg.org/dwc/terms/index.htm> (Darwin Core Terms: A quick reference guide) [↑](#endnote-ref-2)