Humboldt Extension Response

A Miscellany of Fine Discontent

The ALA Data Team 2023-09-14

The Humboldt extension looks like it will be very useful for the ALA’s extended data model (EDM). It gives the EDM a formal, consistent, exchangeable model for describing the sort of datasets (and sub-elements) that the EDM is supposed to service.

# Big Stuff

The ALA is highly reliant on machine processing when it aggregates data. To do this well, we tend to want clearly labelled data that does not require further parsing; particularly if it requires some sort of semi-natural language processing.

Most of the comments in this section are motivated by the desire to structure data in a machine-processable form. This does not diminish the importance of human-readable data since, at the end of the day, inventories are meant to be examined by someone looking for suitable data.

## Inventory Identification and Structure

An *inventoryID* term is needed, so that the inventory can be referenced, if necessary.

Inventories implicitly have a parent-child structure that shadows the event hierarchy in an event core archive. A *parentInventoryID* allows an explicit linking of parent-child inventories, with the same sort of restriction that the child represents a subset of the parent that is present in event hierarchies.

## Multiple Inventory Statements

An event could have multiple inventories. For example, a survey that links insects in ground cover and birds in the canopy layer would be better served by two inventory statements linking the target taxon and habitat. The *parentInventoryID* described above can be used to disambiguate between multiple sub-inventories.

See also, Multiple Inventories for Different Purposes below.

## Lack of Identifiers and Semantic Interoperability

Within the general Darwin Core structure there is a tendency to allow an identifier to be associated with a more general textual term. For example, *datasetName* and *datasetID* from Darwin Core or *measurementType* and *measurementTypeID* from the extended measurement or fact extension. The *measurementType* would contain a string such as “area” and *measurementTypeID* would contain the URI “http://vocab.nerc.ac.uk/collection/P24/current/AREA/”

If a description can be unambiguously associated with a semantic URI, the inventory can be used by machines to aid search and discovery. A resolvable URI also aids user comprehension since it allows a user to locate additional contextual information about a term.

As a general principle, any term that is associated with a formal vocabulary should have an associated ID-term. For example, *identifierTypes* and *identifierTypeID* Similarly, anything that might reference an external definition should also have an associated ID-term that can be used to locate a formal description of the definition. For example, *samplingEffortProtocol* and *samplingEffortProtocolID*.

## Inventories as Generated Description

The extension documentation specifically disavows inferred inventories constructed by aggregating data holdings. The idea appears to be that inventories should reflect statements of intent and statements about how close the survey activities are getting to the intention. Statements of intent and completeness require explicit, supplied data, since they essentially describe the state of mind of the people designing the survey.

This is a good distinction to maintain. However, the concept of an inventory happens to nicely fit the sort of search facets the EDM would like to use – and already uses and generates. There is no point in building a parallel model to keep the Humboldt extension conceptually clean, so it is likely that the EDM will simply reuse the extension model. It would be good to be able to identify inventories by purpose: intention, actuality and description and allow generated description without restrictions. See also, Multiple Inventories for Different Purposes below.

## Multiple Inventories for Different Purposes

In analysing the extension model we identified three possible uses of inventories:

* Statements of intention. This is the scope of what the activity is intended to cover.
* Statements of actuality. This is the scope of what the activity has actually covered, based on comparison with a statement of intention.
* Statements of description. This is the scope of the data collected.

The current Humboldt extension appears to link statements of intention and actuality into a single record. Term names reflect this, with *targetTaxonomicScope*, rather than *taxonomicScope.*

In some cases, it is hard to determine whether the term refers to an intention or an actual result. For example *eventDuration* could mean either the intended duration or the actual duration taken.

We propose a different way of structuring inventories, with *inventoryType* describing the intended purpose of the inventory. This would involve the following changes:

* Terms such as *targetTaxonomicScope* would be renamed *taxonomicScope*. Whether the scope is a target, measure of completeness or a description would be dependent on the inventory type.
* Completeness inventories and description inventories can be linked to target inventories by use of a *targetInventoryID* using the *inventoryID* discussed above.

The changes proposed would allow the presence of descriptive inventories without ambiguity. It would also allow a cleared relationship between intention and actuality; terms such as *hasVouchers* can either represent the intention to collect vouchers or the fact that voucher specimens have been collected.

### Multiple References to Inventories

A survey design that has multiple visits to the same site, with a survey conducted according to a common protocol and intended inventory would seem to be a common occurrence. As far as I can see, the expectation is that the inventory is simply copied with different completeness results. Splitting out intention vs result goes some way towards describing this behaviour, as would allowing inheritance of common inventories, but would require an event structure that groups events by protocol.

Assuming that we didn’t skip over this somewhere, some recommendation on how to structure this sort of thing would be helpful.

## Things that Should be Measurement or Facts

The term i*sVegetationCoverReported* suggests that the cover is reported in *verbatimSiteDescriptions*. Why not use a *dwc:MeasurementOrFact* (MoF) extension and report the metric in a machine-parsable way?

Similarly, *reportedWeather* and *reportedExtremeConditions* will tend to require embedded JSON to properly report. Again, these could be MoFs.

If the values actually apply to the inventory, rather than the event, introducing a *inventoryID* would allow MoFs to be linked to inventories in the same way that occurrences are in the extended measurement or fact extension.

# Small Stuff

A collection of queries, requests for clarity or niggles.

## Including Units in Terms

Including units in term names causes problems. It induces either over precise values or rounding errors. (Eg. is 1 mile 1.609344, 1.6 or 2 kilometres? the choice can lead to quite misleading results if multiple lengths are added together.) It’s also inconsistent with other terms, such as *eventDuration* and *eventDurationUnit*. *geospatialScopeAreaInSquareKilometers* embeds the unit into the concept that is being described. Why not *geospatialScopeArea* and *geospatialScopeAreaUnit*?

## Duplicated Terms

The terms eco:*identifiedBy, dwc:identifiedBy* and *eco:identificationReferences, dwc:identificationReferences* appear to be duplicated. We think that it may be because it is intended for use at a higher level that occurrences however that is not clear from the definition in https://eco.tdwg.org/terms/#identification

## Missing Terms

Is there a reason why there is a *targetLifeStage* but not a *targetSex, targetReproductiveCondition, targetVitality* or other occurrence terms?

## Plural Terms

The terms *protocolNames* and *protocolReferences* are plural but *protocolDescription* is singular. Similarly, there is a *compliationType* but *inventoryTypes* We expect that, in most cases, there will be a single *inventoryType, protocolName* etc. If multiple terms are included, they are separated by vertical bars. As a result, we prefer the singular form.

## Dates and Duration

Should there be a *targetEventDate* or *targetDuration*?

Should there be a *targetSchedule* that describes the level of broad/actual time in the same way as the geospatial scope?

Should there be a *verbatimEventDuration* or *verbatimTemporalScope* for times as reported (eg. “one day”)?

Does the *eventDuration/eventDurationUnit* refer to the expected or actual duration?

Since *dwc:eventDate* explicitly allows for start/end dates and times (eg. 2019-07-02T12:03:00/2019-07-02T12:15:00 has an event duration of 15 minutes, it would seem that the *eventDuration* is the expected duration.

Should *eventDurationUnit* be removed and the duration always be in the form of an ISO 8501 duration?[[1]](#footnote-1)

## Inventory Types

The purpose of inventoryTypes as opposed to compilationType is unclear. This seems to be a common problem, as <http://purl.obolibrary.org/obo/BCO_0000047> contains a similar comment.

## Scope Type

The examples given for *targetTaxonomicScope* are broad taxa at the class level, such as *Aves*. This looks like broad, descriptive information, rather than a specific inventory of expected species. Having a more specific inventory would allow users to infer absences, even if we don’t do so ourselves. A t*axonomicScopeType* of, say, “descriptive” or “normative” would indicate the intended level of the scope.

## Miscellaneous

“This term is only relevant if the *dwc:Ev*ent represents an inventory.“ appears in a number of places. However, the examples tend to attach inventories directly to an event like a survey. What does this mean?

What happens if the *samplingEffortValue* and *samplingEffortUnit* require multiple values? Eg. “2 hours over 100m2”

The terms *protocolNames* and *protocolReferences* are split. Should *samplingEffortProtocol* have the same treatment?

Should *targetCompletenessReported* allow a measure, such as “50%” rather than a vocabulary term?

Terms such as *isSamplingEffortReported* seem to be inferable from other terms. And having *isSamplingEffortReported: true* and *samplingEffortValue* empty makes no sense. Why are these terms needed?

JSON tends to make a value somewhat opaque. If a term contains embedded JSON, then it should be in JSON-LD format, so that the resulting data is comprehensible.

1. https://en.wikipedia.org/wiki/ISO\_8601#Durations [↑](#footnote-ref-1)