2021-09-15 Material Sample Task Group Meeting

Attendees

10AM MDT

Roger Burkhalter, Abby Benson, Sharon Grant, Teresa Mayfield-Meyer, Tim Robertson, Stan Blum, Richard Pyle, Steve Baskauf, Jutta Buschbom, John Wieczorek, WK (?), Stephen Richard

4PM MDT

Roger Burkhalter, Abby Benson, Teresa Mayfield-Meyer, Stan Blum, Mariel Campbell, Michael Hope, Sharon Grant, John Wieczorek, Greg Whitbread, Stephen Richard, Simon Cox

Agenda Items

Tasks from last meeting

Finalize charter

- Charter accepted: https://www.tdwg.org/community/osr/material-sample/

Create mailing list

- Created: dwc-material-sample@lists.tdwg.org

Participate in GitHub Repo discussions

- definitions for the terms indicated in the Primary Deliverable

Recorded Meetings

- There is a folder in Google Drive where I will post meeting recordings

Items for discussion

1. BasisOfRecord
    a. REQUIRED for IPT
    b. Focused discussion - https://github.com/tdwg/material-sample/issues/11#issuecomment-920362964

2021-09-15 Material Sample Task Group Meeting

c. Radical transformation? - probably out of scope, but keep this in our minds
   i. Steve brought up that there are separate issues - one is that we need to define what basisOfRecord means and the second is breaking the star schema and the second one seems out of scope for this task team
   ii. Tim: Material Sample as a core was never moved forward by GBIF because there was never a good justification for creating it that would be different from the occurrences. Darwin Core Archives were not part of Darwin Core.
   iii. Steve: rdf:type, rdf guide is officially adopted by the Darwin Core standard, need to review what's in the guide. https://dwc.tdwg.org/rdf/ 2.3.1.4 A term that we have for describing a type of a thing.
   iv. Rich: if it was in occurrence it would make sense as the "type" of occurrence, but we are also using it for other "types"
   v. Discussion about class hierarchy and ontology
   vi. Sharon: terms in basisOfRecord are conflated (e.g. PreservedSpecimen and FossilSpecimen)
   vii. Jutta: How to relate this to the digital specimen work
   viii. Tim: Seems like we need a Material Core that documents a physical thing that you need an identifier for. Each record represents a preserved specimen/sample with extensions (including maybe an occurrence extension). GBIF is thinking about parsing these types of records by the different types (e.g. a Material Portal, a DNA portal, a Observation portal). We need a richer view of the provenance of each of the record types.
   ix. Rich: Suggests keeping the occurrence core but having a way to document the evidence in some way so we don't end up with pseudoreplication of occurrences
   x. Sharon: Is the occurrenceID the "key" to everything that comes from it (leg, gut contents, DNA)? Or is the occurrenceID for each piece?
   xi. Steve R: Suggestion to focus that the occurrence is an Organism at a place and time and the occurrenceID should be that.
   xii. Michael H: Atlas created their own basisOfRecord for eDNA but when they moved their pipelines to GBIF they lost that term. If we determine basisOfRecord should only be the indication of the evidence how are we going to handle the searching/sorting/filtering that is currently done using this term.
   xiii. John W: We have a short term and a long term solution- short term add "eDNA" as a term. But long term there is an issue that records have multiple types.
   xiv. Michael: The assumption that every record in the Atlas is an occurrence is not accurate. Shared a powerpoint diagram trying to capture what was happening in the MaterialSample github issue that showed that really identification is the key to how basisOfRecord is being used.

2021-09-15 Material Sample Task Group Meeting

basisOfRecord is being used as a proxy for the means of identification (diagram shown starting at 16:46 of recording of 4PM meeting)

xv. Stan: need to document use case

d. How does this affect the prospects for MaterialSample?

2. PreservedSpecimen
    a. Remove “specimen” from the definition
    b. deprecate all xxxSpecimen terms, and instead focus on the distinction between Organism and MaterialSample.
    c. Do we need PreservedSpecimen at all?
        i. At the risk stepping in front the bus, why do we need PreservedSpecimen at all?
        ii. Indeed! I've made my answer to that question pretty clear.
        iii. PreservedSpecimen = MaterialSample + preparation
        iv. On the basisOfRecord issue, I think the problem is that the "pseudo-classes" (my term) of LivingSpecimen, PreservedSpecimen, and FossilSpecimen feel like attempts to overload information into the basisOfRecord values. I see all of these as slight variants of MaterialSample, and should be distinguished from each other through other properties besides basisOfRecord.
3. FossilSpecimen
    a. A preserved specimen that is a fossil.
    b. FossilSpecimen = PreservedSpecimen + (preparation = fossil)
    c. Fossils are not typically something anthropogenically preserved, that is how that class is used, but are instead naturally preserved.
    d. FossilSpecimen is used by collections and researchers to include or exclude within data searches.
    e. Certainly the fact that a physical MaterialSample item represents a fossilized impression of an organism (as opposed to directly preserved tissue from a recently living organism, or a still-living organism) is critical information that should be included within the spectrum of data exchanged using DwC terms. But the question is: is basisOfRecord really the best way to capture this?
4. MaterialSample
    a. A physical result of a sampling (or subsampling) event. In biological collections, the material sample is typically collected, and either preserved or destructively processed.
    b. Instances of MaterialSample are aggregates of physical material that are extracted ("collected") from the natural environment, and held in the custody of humans. These are physical items that may be preserved or destroyed, curated or accessioned, borrowed and loaned, subsampled or aggregated to yield new instances of MaterialSample, and otherwise cared for and/or maintained in some way by humans.

2021-09-15 Material Sample Task Group Meeting

c. A materialSample is an object separated from the material world, intended to be representative of some sampled feature.

Next steps and assignments

1.

Meeting Chat 10AM MDT

From Me to Everyone: 10:01 AM
https://docs.google.com/document/d
https://github.com/tdwg/material-
https://docs.google.com/document/d
From Sharon Grant to Everyone: 10:01 AM
I'm just jumping out and back in again
From Me to Everyone: 10:04 AM
https://github.com/tdwg/material-sample/issues/11#issuecomment-917728044
From Richard Pyle to Everyone: 10:07 AM
I agree with what Steve just said. My comments on the issue were responding to John W.'s comment
From Steve Baskauf to Everyone: 10:11 AM
https://dwc.tdwg.org/rdf/ 2.3.1.4
From John Wieczorek to Everyone: 10:17 AM
I don't see them as multiple functions.
I have Nancy, which makes this insanely difficult.
From Tim Robertson to Everyone: 10:17 AM
The GBIF(!) Event schema is https://rs.gbif.org/core/dwc_event_2016_06_21.xml
From John Wieczorek to Everyone: 10:19 AM
But basisOfRecord was to say what the record was about.
It could be used to share gazetteers, basisOfRecord=Location.
For Event Core, there is only one possible value for basisOfRecord, so it would be superfluous to add it. If we had subtypes of Events (and this has been asked for) then the Event Core too should have a type term.
eventType or basisOfRecord.
From Richard Pyle to Everyone: 10:20 AM
I prefer the sense that you are suggesting here, John -- but the addition of options of "subtypes" of Occurrence (which themselves are a mix of subtypes of MaterialSample and Observation) is where the confusion comes from.
For example, would "PoliticalLocation" and "GeographicLocation" be appropriate values to include for Gazetteer records?
From John Wieczorek to Everyone: 10:20 AM

2021-09-15 Material Sample Task Group Meeting

I don’t see why there should be any confusion. It is the most specific class you can apply.
From Richard Pyle to Everyone: 10:20 AM
Or would it be better to have a term within the Location class that specifies this distinction?
From John Wieczorek to Everyone: 10:21 AM
Subtype or otherwise.
What would the distinction within Location Class signify?
From Richard Pyle to Everyone: 10:22 AM
Hawaii the political entity, vs. Hawaiian Islands the geographic feature
From John Wieczorek to Everyone: 10:23 AM
Sorry Rich, but I don’t see how that is relevant to the issue at hand.
From Richard Pyle to Everyone: 10:25 AM
My point are that there are endless possible subclasses and sub-subclasses and sub-sub-subclasses, all of which could legitimately meet the definition of what the record is about. The distinction between PreservedSpecimen and FossilSpecimen seems much less about the basisOfRecord than about the characteristics of the object the record represents.
From Abby Benson to Everyone: 10:25 AM
I feel like what you outlined here Steve does show that there are two ways to think about basisOfRecord: https://github.com/tdwg/material-sample/issues/11#issuecomment-918534773
Specifically: I think the key thing here is to keep in mind the use cases for why we want to know dwc:basisOfRecord/rdf:type. In my mind there are two big ones:

as a categorization mechanism for searching/sorting/filtering and
to understand the kind of evidence that is available to document an occurrence or determination.
From Richard Pyle to Everyone: 10:26 AM
In other words, I don't see how PreservedSpecimen can be a different basisOfRecord from FossileSpecimen in the same way that a Taxon record is different from a Location record.
From Steve Baskauf to Everyone: 10:29 AM
@rich on a basic level, they are all types of things that a record can be “about”
From Richard Pyle to Everyone: 10:31 AM
Steve, I agree -- but you can say that about almost any term in any class (e.g., "this record is about a specimen collected by Richard Pyle", so basisOfRecord="PreservedSpecimenCollectedByRichardPyle"
Sharon is explaining this better than I did.
From Jutta Buschbom to Everyone: 10:31 AM
https://docs.google.com/document/d/1_q-O4eigdIO9_5BZpZXIIAE5sqi8E9gQREOwJjBjwkk/edit
From John Wieczorek to Everyone: 10:32 AM
@Rich, you could if the vocabulary was “rich”enough.
From Richard Pyle to Everyone: 10:32 AM
:-)
From Steve Baskauf to Everyone: 10:36 AM

2021-09-15 Material Sample Task Group Meeting

The first part of this document really reminds me of what we were trying to do years ago in Darwin-SW. We had a class for “token” similar to what’s described there, and we created a graph do describe the connections.
From Richard Pyle to Everyone: 10:36 AM
👍 @Steve
From Sharon Grant to Everyone: 10:36 AM
yes - that!
From Steve Baskauf to Everyone: 10:38 AM
But what Tim is suggesting completely leaves out born-digital, documented evidence like camera trap images, video, or sounds
From Sharon Grant to Everyone: 10:40 AM
might born digital be a from of preservation?
form
From Me to Everyone: 10:40 AM
+1 Sharon
From Steve Baskauf to Everyone: 10:40 AM
It’s fine if you are running a museum, but not if you are running iNaturalist
From Abby Benson to Everyone: 10:40 AM
That seems like a weird way to think about it for those that are not in the museum world.
From John Wieczorek to Everyone: 10:41 AM
That MaterialCample Core would be only for those types of data.
Others would still exist.
From Sharon Grant to Everyone: 10:45 AM
+1 to that
From Steve Baskauf to Everyone: 10:46 AM
+1 Rich
I think what Tim says makes sense, I’m just not understanding why it has to be limited to physical things. If it were a catalog of Evidence, then born-digital evidence would be included
From Richard Pyle to Everyone: 10:47 AM
I don't think he's saying that GBIF itself will be limited to physical things, but that there would be a "Material" core optimized for those kinds of data(??)
From Abby Benson to Everyone: 10:47 AM
That's what I was understanding too.
From John Wieczorek to Everyone: 10:48 AM
Here is some homework for everyone. What happens to your head if EVERYTHING was based on the equivalent of an Event Core?
From Richard Pyle to Everyone: 10:48 AM
I could get on board with that.
From Steve Baskauf to Everyone: 10:48 AM
@Sharon, that’s basically why we have the Organism class
From Richard Pyle to Everyone: 10:49 AM

2021-09-15 Material Sample Task Group Meeting

@John -- especially if we treat things like "I extracted this tissue sample from this preserved specimen" as an "Event" (as we do in our modelling)
From Steve Baskauf to Everyone: 10:49 AM
@Sharon Again Darwin-SW models this with transitive derivedFrom relationships
From Tim Robertson to Everyone: 10:50 AM
I only focused on Material (i.e. physical stuff) because of the focus of this TG. I have similar monologues about eDNA, and Camera traps etc which have a different set of concerns
From Richard Pyle to Everyone: 10:50 AM
+1 @Tim
From Steve Baskauf to Everyone: 10:51 AM
+1 tim
From Tim Robertson to Everyone: 10:52 AM
John - write this down!
From John Wieczorek to Everyone: 10:53 AM
Will have to come back to the recording for it.
But people are conflating Occurrence.
Happily, Arctos has it correct internally.
From Steve Baskauf to Everyone: 10:53 AM
See Fig. 2 of http://www.semantic-web-journal.net/system/files/swj1213.pdf
From Sharon Grant to Everyone: 10:53 AM
it did
From Tim Robertson to Everyone: 10:55 AM
Conflating occurrence John?
GBIF are beginning to link occurrences together to recreate reality. E.g. a museum specimen, it’s sequence (in 2 databases) and a material citation in a treatment
https://www.gbif.org/occurrence/3346362878/cluster
From John Wieczorek to Everyone: 10:55 AM
Yes. The use cases DO refer to the same Occurrence, but with different MaterialSamples.
From Richard Pyle to Everyone: 10:56 AM
@John -- Exactly! Multiple evidence supporting the same occurrence
From Abby Benson to Me: (Direct Message) 10:56 AM
Just a quick heads up that I have another call at 11 that I have to leave for.
From Tim Robertson to Everyone: 10:56 AM
Occurrence becomes problematic, when subjective information is included, such as the identification (species).
From John Wieczorek to Everyone: 10:57 AM
We have the same thing going on with Organism in lots of use cases.
@Tim, that is why we wanted an Occurrence to be about an Organism at a place and time - the Identification is a separate issue in that case.
From Richard Pyle to Everyone: 10:58 AM
+1 @John
From Sharon Grant to Everyone: 11:00 AM

2021-09-15 Material Sample Task Group Meeting

but the "thing" is a great "thing"
From John Wieczorek to Everyone: 11:01 AM
Occurrence is definitely a central problem. We did it for convenience to flatten what was one of the two most important things to be able to share data about at the time.
The other was Taxon.
From Abby Benson to Everyone: 11:02 AM
I have to run.
From Sharon Grant to Everyone: 11:02 AM
deduplication of evidence
From Jutta Buschbom to Everyone: 11:02 AM
How to organize the a PID is communicated to the different users of an individual/specimen? Is this predominantly a technical or a social-organizational/logistics question?
From John Wieczorek to Everyone: 11:03 AM
PID?
From Jutta Buschbom to Everyone: 11:03 AM
PID: Persistent Identifier
From John Wieczorek to Everyone: 11:03 AM
OK, thank you.
From Steve Baskauf to Everyone: 11:04 AM
Sorry, I also have to go.
From Richard Pyle to Everyone: 11:04 AM
@Tim -- I think one way to maintain DwCA and star schema is to reduce the properties of the Occurrence Core to just those terms organized in the Occurrence class (https://dwc.tdwg.org/terms/#occurrence) and spread out other stuff (Event, Location, Taxon Identification, etc.) into extensions.
From John Wieczorek to Everyone: 11:05 AM
I think the star will be too constraining. We should be able to make greater relationship richness possible. The challenge will likely be for the publisher.
From Sharon Grant to Everyone: 11:06 AM
agreed - it is now limiting
From Richard Pyle to Everyone: 11:08 AM
I think it might be possible to capture everything with only one "hop" if the only Core was ResourceRelationship, and all "ID" terms were represented as RR instances.
From John Wieczorek to Everyone: 11:10 AM
@Rich. True, that makes it completely scalable in terms of complexity, and essentially faking RDF, but that puts the challenges in a different place.
From Richard Pyle to Everyone: 11:11 AM
@John -- exactly. Basically a "fat" triple store with cached extensions.
But agreed on the challenges in a different place thing
From John Wieczorek to Everyone: 11:12 AM
Yes Stan! There are the publishers and the users, and their needs do not always align.
Relevé

2021-09-15 Material Sample Task Group Meeting

From Sharon Grant to Everyone: 11:13 AM
ResourceRelationship is the 1st hop in essence but you don't want to bolt all the terms for each type of evidence into it (It think)
From Richard Pyle to Everyone: 11:13 AM
Honestly, the throttle of progress boils down to available software that people use to manage their data
From Sharon Grant to Everyone: 11:13 AM
maybe - but we can bend software to our will if we know what shape is needed on the other end
From Richard Pyle to Everyone: 11:13 AM
+1 @Sharon
From John Wieczorek to Everyone: 11:16 AM
Teresa, can you get me the link to the recording when done?
From Sharon Grant to Everyone: 11:16 AM
I have to jump out!

Meeting Chat 4PM MDT