TDWG Technical Architecture Group (TAG)
Online meeting
2023-03-13 13:00 UTC

Agenda/Meeting notes:

Participating:

- Steve Baskauf
- Raïssa Meyer
- David Fichtmueller
- Ian Engelbrecht
- Jonathan Rees
- James Macklin
- Kit Lewers
- Camila Plata
- David Shorthouse (late arrival)

Regrets:

- Tim Robertson
- Rod Page

Comments and notes taken during the meeting are in red.
NOTE: please make comments as text directly in the document rather than in marginal comments.

As with previous meetings, Steve will record the meeting for note taking and later viewing.

I. Set time and date for next meeting: 2023-05-08 14:00 UTC.
II. Follow-up on items from the previous meeting
    A. GUID organization meeting (Ian): 2023-03-21 14:00 UTC
        1. Draft meeting document
        2. Background: Agosti et al. https://doi.org/10.3897/rio.8.e97374
        3. Background: EOSC policy document https://op.europa.eu/s/x1W5
    B. Update on standards mapping organizational meeting (David F.)
        1. No meeting yet, but Holly, Mareike and David will meet tomorrow (2023-03-14) to discuss submitting a proposal for a session at TDWG 2023 about Standards Mapping and at that occasion maybe also discuss potential dates for a mapping TG meeting.
        2. James Macklin and Raïssa Meyer communicated interest in participating / contributing
    C. Work on FAQ for newcomers (Kit)
        1. Making a landing page for newcomers to TDWG
        2. Creating an organizational chart
    D. Policy on boolean values: discussion and approval (Steve).
        1. If approved, this would be a TAG technical recommendation, but not a standard.

2. It would go in some "technical recommendations" section (not in the standards page)
3. Draft controlled vocabulary: https://github.com/tdwg/tag/blob/master/build/boolean/index.md
4. Best practices document at https://github.com/tdwg/tag/blob/master/boolean/boolean-best-practices.md Designed to complement Ben's reference document at https://docs.google.com/document/d/14f7H7hSTyiu0LXCJc9nG-uvy3-ElAkUd/edit?usp=sharing&ouid=116355305033345986500&rtpof=true&sd=true
5. Discussion points from last meeting:
   We confirmed that for serializations that support data typing, the values should be data typed as boolean. This view is expressed in the best practice document.
   The GBIF API responds in various ways. dwc:decimalLatitude and dwc:decimalLongitude are expressed as floats. dwc:individualCount returns strings, but that may be to accommodate ranges like "1-2". GBIF expresses their own term "isInCluster" as boolean.
   Comment in doc by Tim Robertson: In which API, please? /occurrence/search returns as an Integer - example https://api.gbif.org/v1/occurrence/4027380301
   The verbatim API will always be strings https://api.gbif.org/v1/occurrence/4027380301/verbatim
   Rod Page says that he favors JSON typing for boolean, integers, and floats. iNaturalist follows JSON typing for integers, floats, and booleans.
   John Wiezorek noted that GBIF follows JSON typing when they present interpreted data over which they control the content and therefore the data type. This is an argument for the verbatim terms where the original data may be presented as strings.
   Ben Norton notes that applications that output data have to cast them as a particular type (e.g., integer). Since people are sloppy you find output both ways (typed or as strings). If we say that you should cast booleans without quotations, we will still get both.
   Steve notes that although we can't enforce RFC 2119 MUST statements, it's a first step towards encouraging people to be more consistent.
   Ben: Awareness will help a lot, people would be less likely to overlook.
   The consensus seems to be that if it's possible to indicate that a value is a boolean, it should be, by whatever means available in a particular serialization system.
6. Translations: https://docs.google.com/spreadsheets/d/1zjq88Y7QNf35u3tAS9ofQEPJeGn9PNiTM0WNVqIu_fo/edit?usp=sharing
7. Comment by David F.: we should include in our guidelines that systems exporting data should be strict in their use of the vocabularies, but that systems importing values should be fault-tolerant and recognize alternative values such as 1 and 0. This

should go in the Best Practices document. It could go in section 7. Steve requested David F. to make a suggestion (pull request).
8. Ian (from chat): David's point about Boolean values and data providers vs data consumers, Jonathan pointed out Postel's Law: https://en.wikipedia.org/wiki/Robustness_principle
9. No one objected, so this was approved.

E. Recommendations for expressing complex values (Ben)
1. Update. He's been working on it. Latimer Core has a consistent way of handling them that is documented clearly. It involves JSON arrays, which are either arrays of values or arrays of JSON objects.
2. Steve: for things that have JSON serializations, you can make them as complicated as you want. What people want is guidance on is situations where you have tabular data. It's straightforward when there is a single value, but if people want multiple values, we are stuck with the "space pipe space" solution, which is the default. But then someone has to parse that out. In the example of dwciri: terms, if the value is an IRI, there shouldn't be any additional cleaning up required. However, if you want multiple values, how do you express that without requiring additional wrangling? For example, do you specify that the value is a JSON array of HTTP IRIs? The other thing that comes up (e.g. in Humboldt Extension), when you have quantities that require two properties to fully describe (e.g. magnitude and unit), you're fine if there are single values, but you can't put multiple pipe-separated values in the two fields because you don't know which value goes with which unit. People want a solution for that. At what point do things get so complicated that you just have to have a relational database and forget about trying to cram things in a single cell of a flat table.
3. Ben N.: That's the two types of arrays. The simple arrays are just a list of values (like a pipe separated list of values). Then there's the more complex arrays, which is a JSON array of objects (for example "common name":value, "language":value, "primary":value). So the question is "is that a class"? If it's a class, the syntax is "hasVernacularName" and "VernacularName" would be a class. If it's not a class, there has to be a way to make a distinction. What is the situation where you have an array of objects where it is not a class? Is that possible?
4. Steve: Darwin Core has three main categories of properties: properties for which a string value completely denotes the object (example: decimalLatitude), properties that link to an entity that's the instance of some class (example: recordedBy whose value is an Agent; essentially an object property sensu OWL) and that links to some thing that you might describe in another table, properties that have controlled values that link to SKOS concepts (example: establishmentMeans). The last category could be considered to be a special case of the second category, but you are trying to direct people to a particular controlled value, not connecting them via a link to another entity.

5. David S in chat: A risk here to get wildly out of scope (eg n-tuples). I recommend keeping clean and simple w/ JSON arrays and arrays of objects.
6. Ben: the terms that have an array of objects have "hasX", X isn't necessarily a class, but it's something that has multiple attributes. Walking through examples are very helpful.
7. David Shorthouse in the chat: { length: { value: 12, unit: “cm” } }
8. Ben: That's an object but not an array since you don't have multiple lengths. In that case you'd need to have "hasLength".
9. David F in chat: What about author teams and specifying the order of the authors. This is not a property of the class, but of this particular configuration of the particular publication. Would this be an example that you were asking for or would that be a different problem?

F. Report on efforts to document the status of prior and current standards (Kit)
   1. She's reaching out to publishers for some of the written prior standards (in or out of print) to see if we can digitize and post them.
   2. She's checking with Theary Ung whether some of the older geographic standards can be subsumed by the new work the Geoscheme's group is doing.
   3. Draft survey to find out who's using the standards (to be promoted at the intro session of the next annual meeting): https://docs.google.com/forms/d/1JuJTBA-Dkz2L75Eu3XsaBWnXbuZ1NMu7IXiHI096gCc/edit – Link to the field, should be available to anyone with the link :)
   4. How are they using the standards
   5. Additional comments
   6. James: do you know of others who are using the standard (so that we can ask them to participate in the survey) outside of our community.
   7. Kit is going to do some web scraping of Semantic Scholar to find out who's citing them.
   8. A general field for additional comments (in particular "how do you use the standards") might be nice.

III. New items

A. Formalizing a process for giving feedback and suggesting changes to existing standards that don't have a Maintenance Group and issue tracker (e.g. SDS, VMS, GUID AS, Tapir).
   1. The TDWG Process document says that "An Interest Group is also responsible for maintaining the products of its past Task Groups", but there may not actually be a functioning IG any more. In the case of the SDS and VMS, the Process IG sponsored the TG. The Process IG does exist, but isn't active and its charter looks really out of date. There is a GitHub repo and issue tracker.
   2. It seems like the process should be detailed in a similar manner to how the VMS describes the process for maintaining vocabularies.
   3. James: cases where there's no obvious maintainer should default back to the TAG. The TAG could recommend to the Executive that if a standard needs updating and there isn't a maintenance group, that a

group be formed or if the changes are small, the TAG itself could handle it to avoid the usual problem of having to identify leadership to take on the task.

4. Steve: he considered the TAG as an option, but in the past the TAG also wasn't always active, which would result in the same problem. Unlike the VMS, which clearly lays out the process of vocabulary maintenance, there is no clearly defined process for maintaining other standards (e.g. who manages the process, what exactly is supposed to happen). In the VMS process, the Maintenance Group stands in the role of Review Manager. That presupposes that the MG is composed of a group of experts in the subject (vs. a single person).
5. Steve: he sees the TAG's responsibility is to "lubricate the wheels" of the TDWG process. So if it sees a need, it could either push the relevant Interest Group to take action, or request that the Executive Committee take action to form a group (or do it themselves).
6. Raïssa: One problem is that TAG members might not have expertise themselves, but they could have knowledge of who those experts are and reach out to them (directly or through the Exec.). That would avoid putting too much pressure on the TAG itself to maintain something that they don't have background on. Steve also noted that the level of commitment on the part of the TAG did not really include this level of work (vs. meeting bi-monthly in an advisory capacity).
7. Steve: there has been some discussion of revising the By-Laws (the "Process" document), which carries a very heavy load with few details.
8. Ian: this relates to the bigger question of people's overall involvement in TDWG (participation, amount of time they have to give). There is often a lot of interest and discussion, but then the activity fades away. How can we "enroll" people to become more involved in TDWG activities and these kind of processes generally? In his organization they have consultants who help break down barriers and make people more interconnected as a community. It does help build relationships and interest.
9. Kit: is there someone on the Exec whose responsibility is to keep an eye on standards to see if they need updating? Could the be an annual or biannual review? Steve: he thinks that's more or less the TAG's responsibility: keep our fingers on the pulse. Because the TAG is diverse and involves people from all parts of TDWG, hopefully someone would be aware of the problems.
10. Steve: maybe make sure that each standard has an issue tracker (perhaps monitored by the TAG). If no issues come up, then we could assume no action needs to be taken.

B. Best practices for borrowing terms from non-TDWG vocabularies.
    1. Recommended course of action from last meeting's notes:

a) Seed the TAG Slack with the Latimer Core criterion for deciding whether to borrow or mint and continue the discussion offline. If something concrete happens there between now and the MaterialSample TG meeting next week, we can take that to them as a recommendation.
b) TAG members who have a long view of how this issue has played out in the past should attend the MaterialSample meeting and contribute to the discussion.
c) As the MaterialSample group decides on this particular proposal, they should also articulate the "short set of conditions" they used to decide their recommendations. That can be developed into a precedent/policy.

2. Is this something where we want to write up a best practices recommendation doc to put in the same place where we are putting the boolean recommendations.
3. This is no longer a time-sensitive item since the Material Sample group has come up with its own solution.
4. Steve: is this a situation where we take whatever wisdom we've arrived at and write up a TAG technical recommendation similar to the other ones we are working on? We could also push this off on a future "mapping" task group.
5. James in chat: Other data standard bodies should have addressed this... maybe they have recommendations?
6. Jonathan: Yes, people have wrestled with these issues for years. They come up with solutions they come up with. It may just be a question of what community you are in. There's no good or "right" answer. What you want is whatever is most harmonious for this community.
7. Steve: The solution that most people seemed comfortable with was: if "same as" relationship: borrow, if not: mint. Audubon Core didn't really do that and therefore has "usage guidelines" for a number of borrowed terms. But that was a decision made at that time.
8. We can continue discussion of this at the next meeting.

IV. Any additional announcements or new business.
   A. none

V. Action items for next meeting (or before):
   A. Steve: publish booleans documents somewhere on the TDWG website.
   B. Steve: check for repositories and issue trackers to see if there is an obvious one for each standard.
   C. Camilla: volunteered to do the Spanish translations for the Boolean vocabulary.
   D. Ben: finish draft of multiple values document