**Environmental Health Language Collaborative**

*Use Case – Data Discovery*

November 16, 2021

**Attendees:**

Amanda Burns, U of Illinois

Anna Maria Masci, NIEHS

Antony Williams, EPA

Arcot Rajasekar, UNC

Canden Byrd, ICF

Cataia Ives, RTI International

Charles Schmitt, NIEHS

Cynthia Grondin, NCSU

Elizabeth Hodgson, ILS

Evan Bolton, NCBI

Knut Erik Tollefsen, NIVA

Kristan Markey, EPA

Maria Shatz, NIEHS

Michele Taylor, EPA

Michelle Angrish, EPA

Mike Conway, NIEHS

Natalia Vinas, Army ERDC

Qianjin Zhang, U of Iowa

Ronglin Wang, EPA

Sean Watford, EPA

Shannon Bell, ILS

Stephanie Holmgren, NIEHS

**Agenda**

* Michelle: Today we will more clearly define the objective of our use case and add clarity around what needs to be developed. As a refresher, review the attachment and please come to the table prepared to discuss some of the main talking points from the previous meeting including (included in the e-mail attachment). I would like you all to think about these in terms of challenge = possibilities! But we have to be SMART.

**Update and review of use case (10 min)**

* Michelle reviewed the workshop use case work-a-thon summary she attached to her preliminary agent e-mail.
  + Qianjin (from Chat): One of challenges we recently observe is that citation sources (i.e., Web of Science) might not capture associated datasets when indexing articles even though the associated datasets are displayed in the publisher's website. There might be a communication problem between different vendors.

**Open discussion on defining the objective and deliverables keeping SMART goals in mind (25 min)**

* Michelle: I want us to think about the next steps in moving this use case forward. We need to identify the objective to work forward, assemble the resources and core trainings to support finding and creating structured data.
  + Charles: We want to get recommendations together for January 2023 when the NIH Data Sharing Policy comes into place. As a product, I am interested if we can put together a recommendation for metadata that can be captured on toxicology data.
    - Kristan: Are you saying metadata to identify what is embedded in the actual data source? Or the field and field names, etc.?
    - Charles: I assume the metadata describing the study (e.g., chemical, species, cell lines, etc.) so people can search across repositories.
    - Shannon (from Chat): That is something we have been working on- defining the minimal information needed for understanding an experiment.
      * Michelle (from Chat): What is your timeline?
      * Shannon (from Chat): @Michelle We are iterative design.
    - Anna Maria: You want to be sure you are capturing what is needed to search for data. You need to consider how the data were produced.
    - Natalia: I like that idea. What are the minimal requirements so people can really work with the experiment? It will depend on your model.
    - Kristan: That would be more of the actual data itself versus tagging it as metadata describing the dataset.
    - Maria (from Chat): One metadata element that keeps coming up is units. Perhaps, creating a list of recommended units for all relevant measurements and socializing it with professional societies, repositories and journals can be a practical first step?
    - Shannon: What would the outcome be?
    - Michelle: We also did not talk about the target audience. The target audience would be large.
  + Michelle: I like this idea for an objective. We discussed this quite a bit during previous workshops, and it keeps coming up. We need to figure out how deep we want to go, how we develop a standard around a minimal criterion but keep it informative and efficient, so it is a format that is adoptable and doable. I know there are some templates out there. What have we learned from these templates?
    - Charles: There are many starting points. Maybe we start by looking at OECD.
    - Kristan: Looking at the different approaches, they all have strengths and limitations.
    - Stephanie (from Chat): <https://github.com/zacharewskilab/MIATE>
* Ronglin: Is one of the objectives to consider human and machine accessibility issues for the datasets?
  + Shannon (from Chat): That was def covered in the discussion- both human and machine friendly.
  + Michelle: That is on our radar. When we discuss minimal criteria, there will be minimal criteria requirements. We need to define those - human and machine accessible, data formats, etc. Is this objective in our power to accomplish? We might have to break off into subgroups to achieve this goal.
  + Qianjin (from Chat): Minimal metadata requirement for datasets and documentation and data dictionary for human to use the datasets?
* Mike (from Chat): We need to think about standard ways to serialize minimum data (e.g., via a spreadsheet, via json-ld).
* Evan: Would you not also want to recommend terminologies to be used since some already exist? Can we get the journals to agree on the terminology to use?
  + Michelle: That might also fit under the requirements.
* Shannon: One thing raised was the enforcement and what that would look like. When we get the requirements, we should look back at "Achievable" and what an ideal situation would look like. Breaking into two parts would be beneficial.
* Michelle: So, a deliverable could be a guidance document.
  + Shannon (from Chat): NTP/CEBS has been working on data dictionaries that may end up being useful? I am sure others have them as well.
* Michelle: So, we would have the table itself and then several deliverables within that - guidance, data dictionaries, training, resources.
  + Antony (from Chat): What about some form of standard reporting tool(s)? As an example of a recent effort: <https://nontargetedanalysis.org/srt/>.
* Michelle: Stephanie, anything else we should consider?
  + Stephanie: There have been several worthwhile ideas put forward. I think we should focus on where the practical first step is to work on. The guidance document will be later. What should come first? Your idea of having subgroups could be another approach.
  + Michelle: Charles, what types of studies do we want to cover? Human, animal, in vitro, in silico?
  + Charles: I would say all.
  + Natalia: I agree.
  + Charles: Maybe we reach out to modeling groups to determine their high priority datasets.
* Mike (from Chat): +1 to this notion, there was some discussion in the DataONE community on this issue, using schema.org and json-ld to capture these relationships.
* Knut (from Chat): Sorry- must leave slightly early today. Some thoughts for the records/discussion: 1) would one of the objective also include developing review/position/research papers?, 2) Would it be beneficial to identify a few core use cases spanning a smaller part of the larger data domain (human to health, different levels of biological organization). These "training sets" can then be used to test applicability to the larger data/knowledge domain so not having to cover all combos at the same time.

**Discussion Review – adding clarity around what need to be developed (15 min)**

* Objective: Develop minimum criteria tables for study reporting.
  + Triangulate between our group
  + Identify external resources (folks who have already worked to develop these types of tables) (Kristan: involve OECD developers)
    - Shannon (from Chat): <https://www.oecd.org/ehs/templates/> this?
    - Consider formats/compatibility/usability/accessibility
  + **Action Items:** Collaborative space for this group to work (potential: Teams, GitHub, Slack, Google Drive, Box, eSharePoint)
    - Add expert areas
    - Add availability
  + Break out into smaller groups for covering the objective and deliverables
  + Define deliverables
    - Guidance document or white paper
    - Data dictionaries
    - Training/resources
    - Kristen: There is metadata, format, and how you link some of the data in the format to other resources. That could be a possible deliverable. What semantic resources could be utilized?

**Out scheduling and wrap-up**

* Michelle: How often should we meet? Any objections to monthly/every 6 weeks?
  + Kristan: I think every 6 weeks is a good starting point.
  + **Action Item:** Michelle: We can send around a Doodle poll to fill out availability for 2022 (will consider European time).
* **Action Item:** Michelle: We will also have folks assign themselves to identify subgroups based on expertise (will be in a table in our collaborative space).