Unlocking semantic phenotypes for the masses: a litany of opportunities

Matt Yoder - Bonn, 2019 Semantic Data Models in Anatomy

View in browser.

This talk was written in **impress.js source**. Source is at **https://github.com/mjy/presentations/tree/master/2019/SemanticPhenotypeModelling**

Other talks from the workshop are collected at https://www.researchgate.net/project/Workshop-Semantic-Data-Models-in-Anatomy

A list

- 10 things that have consequences for how we model anatomy
- Focus on requirements of taxonomy
- Not all doom and gloom

0 - An example item

0 - ... and its Consequences

- One
- Two
- * A consequence that requires no action
- ? A (more) poorly thought out point

1 - Taxonomists have published in *natural language* for over 200 years

1 - Consequences

- * No model is needed?
- A model must support what they do (diagnose species)
- A model must let taxonomists flow naturally from NL observations to a formal representation
- A majority(?) of statements in the model will be lossy

2 - Life is complex

2 - Consequences

- Giant empty matrix, with few links
- Linking nodes must be very carefully thought out
- Models must isolate labels from concepts
- RDF labels are not enough to identify "the same" nodes in disparate named graphs
- Model must support evolving refinement

3 - Life is observed once, described, then ignored

Most anatomical descriptions will never be revised

3 - Consequences

- Integrating previously described species will require NL processing
    - Let's be real- for most species descriptions will not be redone in a "native" semantic format, we simply don't have the time/resources
- ? While our model's semantics must evolve, the statements/observations behind them won't
- ? Our model must have "versioning" to reference the NL algorithm that were used
- Our model must differentiate NL processed statements from "native" statements

4 - Taxonomists present *species* descriptions

4 - Consequences

- Is it important to provide a model that does more than what taxonomists want to do?
- Few, though growing numbers of taxonomists, uniquely identify the specimens in their study
    - If we can't identify specimens, how are we going to identify their parts
- An *instance anatomy* sensu our workshop has never been published

5 - Reference ontologies (for gross anatomy) do not exist for most of life

5 - Consequences

- ? Instance anatomies can't be merged at finer levels of granularity
- ? Search/filter will need to be done on values, and therefor be variously unsatisfying
- We must tackle the problem from top down and bottom up

6 - Converting human NL to a model is lossy process

6 - Consequences

- The model should emphasize *minimizing* loss of meaning
- No one model will fix this issue
- Similarly, conversion between representation models will be lossy, also suggesting minimizing loss is a goal- how to ensure this with model semantics?

7 - Humans can't agree

7 - Consequences

- Semantics need to be fuzzy enough to draw conclusions across independent observations
- It is unlikely we can have 1 graph of observations (e.g. instance anatomy) per entity being described
- Merging/syncing data from the same, or different models remains, as always (sigh), the hardest problem

8 - Model organisms are described differently

8 - Consequences

- Models must account from difference from "normal"/"wild type" type statements
- We must work hard to escape from this relative approach less it persist into a more general usage

9 - All models need interfaces

9 - Consequences

- Interfaces bias what and how models get used
- Semantic models could be completed buried behind the symbolic representations that are used to capture data
- Should our model be built to pre-adapt attributes/properties to the "visual" interface that will capture their instances?

10 - People want to use semantic phenotypes for AI, VR, and other buzzworthy things

10 - Consequences

- Data may need specific attributes to make them useful for AI and other approaches
  - We should talk with Jim/the **SCATE** project
- We need 3D coordinates for anatomy terms
  - See our **vronto** project

11 - ? Data are always generated with a purpose in mind

11 - Consequences

- ? Published descriptions are not "inaccurate" (sensu our discussion in the workshop), they have fulfilled their purpose (and been accepted by a community of peers)
- ? Even if we model data, we can't escape the baggage that is its original purpose?
- ? Even in a universal model, some argue that data derived for one purpose are not suitable for another purpose

12 - URIs/IRIs

12 - Consequences

- ? It is hard to *maintain* and generate *resolvable* URIs at scale
    - ? A universal model requires unique ids, this comes with significant issues such as services that ensure minted URIs are indeed unique
- It is almost certain that we need centrally managed data "lakes"/"oceans"/repositories that our data can find their way to

Conclusion

- It is important that we look hard at our underlying premises
  - Do they reflect reality?
  - Do they reflect how work in biodiversity is actually done?
  - Do they "scale to biodiversity"?
  - Do they reflect realistic applications of technology?
- Keep data representation and data production issues isolated
- A litany of opportunity requires a pluralist approach to encouraging the use of semantics