Social and Semantic Computing in Support of Citizen Science

Joel Sachs and Tim Finin

Computer Science and Electrical Engineering
University of Maryland, Baltimore County
{jsachs, finin}@cs.umbc.edu

Abstract. We describe our ongoing work on using social media as a platform for citizen science. Building on our previous work of facilitating citizen science observations, and using RDF to integrate them with existing biodiversity knowledge, we are currently building Facebook Apps that will enable the reporting of observations, as well as the browsing and tagging of existing observations. The tagging capability serves two main purposes. First, it permits (and, we hope, encourages) multi-stage crowdsourcing for image identification. Second, it serves as a driver of ontology evolution, and permits experiments on potential working relationships between expert-engineered ontologies, and tag-based folksonomies.

Keywords: Semantic Web, Social computing, Biodiversity informatics, Citizen science, Collaborative ontology development

1 Introduction

Species' geographic distributions and phenology (the timing of life cycle events) are changing rapidly in response to climate change, new pathways of migration, and other factors. Observations by amateurs are often crucial in understanding this response. Our previous work [1] investigated social computing mechanisms for publishing citizen science data on the semantic web, where it can be integrated with other sources of biodiversity and biocomplextity data (e.g., range maps, food webs, evolutionary and taxonomic trees; conservation and invasiveness status, etc.) already exposed as RDF. The system concept we envision is a "global human sensor net" – a data stream that can be mined for species of interest (e.g., invasive, threatened, etc.) and anomalies (e.g., species out of their known range.); and which supports drilling down on observations to see what relevant related data (e.g., genomic, behavioral, etc.) already exists in our knowledge base or on the Semantic Web.

We are currently building Facebook Apps that will enable the reporting of observations, as well as the browsing and tagging of existing observations. The tagging capability serves two main purposes. First, it permits (and, we hope, encourages) multi-stage crowdsourcing for image identification. Second, it serves as a driver of ontology evolution, and permits experiments on potential working relationships between expert-engineered ontologies, and tag-based folksonomies.

The motivation behind using Facebook as the platform is to expose observing and tagging activity in users' news feeds, thus facilitating conversation around observa-

tional events. Observations and their tags are stored in Google Fusion Tables, which, in turn, are used to drive RDF representations. There is some contention over what RDF representations of ecological observations, and the ontologies behind them, should look like, and one of our desired and expected contributions are RDF representations of biodiversity data demonstrated to satisfy typical citizen science use cases. Thus, although our Facebook app is itself small in conceptual scope, it serves as a microcosm for a number of design decisions facing the semantic web for biodiversity informatics.

2 Related Work

2.1 Ontologies for Biodiversity

Occurrence Data

The central unit of biodiversity informatics is the *occurrence*, the observed presence of an organism at a particular place and time. Chapman [2] provides an excellent overview of the uses of primary biodiversity (i.e. occurrence data), include building range maps, niche modeling, and gap analysis. The exchange standard for biodiversity occurrence data is Darwin Core, a collection of several hundred terms for describing properties of an occurrence. An important aspect of Darwin Core is that it does not distinguish between data and metadata, so *identifiedBy*, *scientificNameID*, *verbatimCoordinates*, and *eventTime* are all simply properties of the occurrence. There are no mandatory fields.

The TDWG 2010 Annual Meeting in Woods Hole sponsored a day-long bioblitz with the aims of demonstrating TDWG standards in action, and evaluating their potential for uptake and use in real world, citizen science events to serve as a testbed for experiments in social and semantic computing for citizen science. After the bioblitz, a number of long discussions broke out on the tdwg-content regarding the appropriate direction for Darwin Core and related standards [3]. These included questions of normalization, dealing with multiple identifications (both competing and reinforcing), dealing with introduced and cultivated species, GUIDs for taxon concepts, and the meaning of "occurrence". No consensus has been reached, and there are two fairly well developed approaches on the table that we are aware of: that of deVries [4]; and that of Baskauf/Webb [5]. In addition, there is our own representation, which we used to represent the bioblitz data [6]; this serves more to explore the limits of what is possible with a casual approach to knowledge representation, than it seeks to compete with the other two, more principled, approaches, as a possible standard.

Observational Data

We are often interested in knowing more than whether or not a species is present at a location. We may want to know quantitative measurements of physical charecteristics, or qualitative descriptions of phenophase, or descriptions of ecological interactions. Biologists' field note books are notoriously idiosyncratic, and there are a number of proposed models, and, more recently, ontologies, that have been proposed to accommodate the full diversity of observational practice. These include OBOE (the

Extensible Observation Ontology), Prometheus, Delta, and EQ (the Entity-Quaality Model), all of which decompose the observational process in slightly different ways.

2.3 Collaborative Ontology Engineering

Web 2.0 was interpreted in a number of ways, in regards its relationship to the semantic web. For much of 2005 and 2006, it was in vogue to refer to Web 2.0 as the *lowercase semantic web.* This term conflated a number of things: the success of free-tagging to attach keywords to non-text objects; the folksonomies that resulted from said tagging; the embedding of semantics within HTML; and the notion that semantics is best built from the bottom up, rather from the top down.

Almost immediately, upper case Semantic Web researchers sought ways to harness the obvious power of socially created semantics to drive the "real" semantic web. Conceptually, we can divide the resulting collaborative knowledge engineering efforts into two categories: those in which the participants know that they are collaboratively building ontologies, and those in which the the ontologies (or other KR artifacts) emerge from the behaviour of users seemingly engaged in other, non-KR, activities. A large literature exists in both areas, and Angeletou et al. [7] provide a useful guide. Here, we describe only the work most relevant to our own.

Deliberate KR

Siorpaes and Hepp [8] describe a wiki-based approach to marrying ontology engineering and collective intelligence. They contrast *engineering-oriented* ontology design (by far the dominant paradigm) with a *community-oriented* approach, and motivate the need for the latter by listing three main advantages: inefficiency of the engineering oriented approach at keeping up with changing conceptual dynamics; distribution of the KR burden; and higher likelihood of community buy in.

KR as an artifact of user behaviour

Passant describes a system [9], in which tags are associated to concepts in an ontology. If a tag can't be mapped into the ontology, the knowledge engineer takes this as a clue that the ontology needs revision. Thus the traditional domain expert/knowledge engineer partnership is preserved, but with the domain expert role being replaced by the collective wisdom of the community. Passant's focus was information retrieval, where the only reasoning is using subsumption hierarchies to expand the scope of a query, but the principle should apply to other reasoning tasks as well.

Pitts [10] noted that tagging appears to have hit an innovation plateau because it is difficult for users to add more than shallow, impressionistic meaning to a subject, and worked on two projects, Memecat and Listgasm, to encourage meaningful tagging. In order to add the third "predicate" dimension to the tagging of a subject, he provided cues as to what the tagging context is when a user enters tags.

3 Facebook as a Platform for Citizen Science

Facebook may be an excellent platform for citizen science. Incorporating observational events in a user's news feed serves to expose the event to many potentially interested parties, fosters discussion around the event, and promotes discovery of the reporting tool, thereby resulting in more observations. We describe two apps that we are currently developing. One, iDentify, enables multi-stage crowdsourcing of images. The other, iPhenology, enables the reporting of phonological observations. Both apps result in data being published in RDF, and provide us the opportunity to experiment with RDF design patterns for representing biodiversity information. Using tag clouds to annotate the images also enables us to experiment with relating folksonomies and ontologies.

3.1 iDentify

After the bioblitz held at TDWG 2010, we had several hundred unidentified photos. To address this, a webpage (the Taxonomizer) was set up which presented users with an unidentified image, and requested classification. But this resulted in very few new identifications. Two issues with Taxonomizer were i) no one knew about it; and ii) potential users had to sit through many images that they did not recognize before coming to images that they did. iDentify addresses this by allowing identification to occur in stages. If an image is tagged "butterfly", for example, the butterfly experts can look at it to classify it further. Experts can learn about the image either by seeing a post on their wall that says "Your friend has just tagged XYZ 'butterfly'", or by adjusting their settings to show only pictures tagged butterfly.

3.2 iPhenology

Phenology is the study of the timing of life cycle events. For plants, these include first flower, first leaf, leaf senescence, etc. For animals, these include nest building, mating, migration, food gathering, etc. Two major citizen science initiatives in the U.S. capture phonological data: the National Phenology Network, and Project BudBurst. They each provide a controlled vocabulary for describing phenological events. A few things worth mentioning are: i) these two vocabularies use identical terms to mean slightly different things; ii) each vocabulary uses terms not in the other; iii) the NPN vocabulary was revised in the Spring of 2011, illustrating that it is still in flux. In addition to the evolving "standard" phenophase vocabularies, there is rich scope for unexpected, unconventional phenophase description. For example, there is growing interest in tapping into aboriginal knowledge to understand the Boreal Forest's response to climate change, and aboriginal terminology is likely to differ considerably from the terms already defined. Thus the iPhenology app we are developing seeds a tag cloud with terms from these vocabularies, prompts users to select terms from the cloud, and also to free tag where appropriate.

4 Representing the Data in RDF

Darwin Core

One hypothesis is that ontologies for the artifacts of human behaviour should be less constrained than ontologies for the natural world. So, in representing Darwin Core in RDF, we are not concerned with relating the concepts of occurrence, event, location, specimen, etc., through the use of intricate collections of *is_a*, *has_a*, and *part_of,* relations; and heavy use of domain and range constraints, and functional and inverse functional properties. Rather, we see the appropriate place for such ontologies as being the controlled vocabularies that are used as the objects of Darwin Core (DwC) predicates, (rather than for relating DwC predicates themselves). In other words, we see more value in using ontologies to model biodiversity ("tree has_part fruit", "green is_a colour", "human is_a ape", etc.), than in using them to model biodiversity informatics "observation has_part individual", "individual has_part taxon concept", etc.). Therefore, rather than defining an occurrence semantically - for example as the intersection of an event, an individual organism, and an observer - we consider it purely syntactically, as a tuple of time, location, and individual, together with some optional properties.

Flat vs. Hierarchical Ontologies

The notion persists that anything flat is not a "real" ontology, or somehow not semantic. But semantics accrue via human agreement, and do not depend on the topology of the representation. Consider, for example, the following two representations of an occurrence. In the first, *scientificName* is a property of *Occurrence*, while in the second it is a property of *Identification*, which is itself a property of *Individual*, with *Individual* being a property of *Occurrence*.

```
<Occurrence>
  <scientificName>mus musculus</scientificName>
  <individualID>145</individualID>
<Occurrence>


  vs.
<Occurrence>
  <hasIndividual rdf:resource="http://myMuseum.org/specimens?id-145"> /
</Occurrence>

<Individual rdf:about="http://myMuseum.org/specimens?id-145">
  <hasIdentification
rdf:about="http://myMuseum.org/identifications?id=CD/>
</Identification>

<Identification rdf:about="http://myMuseum.org/identifications?id=CD">
  <scientificName>mus musculus</scientificName>
</Identification>
```

The semantics of the above are the same, namely: "There's a thing in the museum that someone thinks is a mouse." We know that, in a sense, semantics transcends worldviews; otherwise people would never understand each other. Often, with no loss

of semantics, the model can be left out of the representation; data can be represented simply as a series of key-value pairs, and then the consumers can ingest the data into their own models.

For representing phenophases, we forgo (for now) the observational ontologies mentioned in Section 2, and instead make use of two terms from the Darwin Core measurement class: *measurementType*, and *measurementValue*. This allows us to embed the phenophase observation within a Darwin Core occurrence record as, e.g.

```
DwC:measurementType phenophase
DwC:meaurementValue first_flower
```

To the extent possible, we represent competency questions as sparql queries (see, e.g., 11), and use these to evaluate our approach.

5 Conclusions

Our current development effort is aimed at answering three questions: Can appropriate tag-cloud interfaces serve as feedback mechanisms for ontologies, and be used to propose new terms?; Can simple RDF representations of biodiversity data support citizen science use cases?; and Is Facebook a good platform for citizen science? We invite comments on our approach, suggestions for further use cases.

6 References

1. Andriy Parafiynyk, Cynthia Parr, Joel Sachs and Tim Finin, Adding Semantics to Social Websites for Citizen Science, Proceedings of the Workshop on Semantic e-Science, AAAI Press, June, 2007.http://ebiquity.umbc.edu/p/365
2. "*Uses of Primary Species-Occurrence Data*". Report for the Global Biodiversity Information Facility 2005. 111pp. (2005) Copenhagen: GBIF.
3. http://lists.tdwg.org/pipermail/tdwg-content/2010-October/thread.html
4. http://www.taxonconcept.org/
5. http://code.google.com/p/darwin-sw/
6. http://www.cs.umbc.edu/~jsachs/occurrences/TechnoBioblitzOccurrences.rdf
7. Angeletou, S., Sabou, M., Specia, L., Motta, E., (2007) Bridging the Gap Between Folkso- nomies and the Semantic Web: An Experience Report. Workshop: Bridging the Gap be- tween Semantic Web and Web 2.0, European Semantic Web Conference.
8. Katharina Siorpaes and Martin Hepp: myOntology: The Marriage of Collective Intelligence and Ontology Engineering, in Proceedings of the Workshop Bridging the Gap between Semantic Web and Web 2.0 at the ESWC 2007, Innsbruck, Austria, June 7, 2007.
9. Alexandre Passant, Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs: Theoretical background and corporate use-case, Proceedings of the International Conference on Weblogs and Social Media, AAAI Press, March, 2007.
10. Blog post: http://www.semanticwave.com/blog/archives/2008_01.tt
11. http://www.cs.umbc.edu/~jsachs/occurrences/queries/sample.txt