Adding Semantics to Social Websites for Citizen Science

Andriy Parafiynyk, Cynthia Sims Parr, Joel Sachs, Tim Finin

Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250 USA
{andr1, csparr, jsachs, finin}@umbc.edu

Abstract

While efforts are underway to represent existing ecological databases semantically, so that they may be intelligently queried and integrated by agents, less attention has been paid to 1) rapidly changing datastreams, and 2) unstructured data from amateur observers. We describe the development of two tools that interact with popular social websites as a means to generate and take advantage of semantic web content for citizen science. Splickr, a website, interacts with the Flickr and Yahoo maps APIs to provide a convenient way of browsing and querying Flickr's geotagged photos. SPOTter, a Firefox plug-in, is an aid to semantic ecoblogging. Both tools generate RDF based on rich OWL ontologies. This approach has wide applicability both in and outside science.

1. Introduction

Scientists require large amounts of both experimental and observational data. While efforts are underway to represent existing databases semantically, so that they may be intelligently queried and integrated by agents, less attention has been paid to 1) rapidly changing datastreams, and 2) unstructured data from amateur observers. The SPIRE project (Semantic Prototypes in Research Ecoinformatics - http://spire.umbc.edu) is addressing this gap for the domain of invasive species science. We recently have begun to develop two tools that interact with popular social websites as a means to generate and take advantage of semantic web content for citizen science. In particular, we built Splickr, based on the Flickr API and developed SPOTter, a Firefox plug-in for semantic ecoblogging. These tools generate RDF based on rich OWL ontologies. In this paper we describe those applications and the larger vision for our future semantic web research.

1.1 Invasive Species

Species that are introduced into ecosystems in which they are not aboriginal are classified as non-native or exotic. Invasives are the small subset of non-native organisms that, through uncontrolled spreading, damage or displacement of native species, disrupt ecological processes and productivity, or threaten human health. Famous invasives in North America include zebra mussels, the Asian longhorn beetle, West Nile Virus, and Chinese snakehead fish; not so famous invasives include sudden oak death, leafy spurge, and innumerable algae. Several thousand weeds, crop pests, plant diseases, disease-vector insects, exotic predators, etc. are of active policy concern in the U.S (Pimental et al. 2000)

Invasive species are thought to be one of the two most important causes of declines and extinction of rare species, and cost the U.S. economy over $138 billion per year (Pimental et al. 2000). Generally, once an invasive species establishes itself in a new environment, it is very hard to eradicate. Early detection and correct classification is therefore critical. It is important to engage as many people as possible to monitor ecosystems and help them promptly report new discoveries, and that includes experienced scientists as well as amateurs. Most new invasive species are first reported by amateurs, and there are several federal, state and local projects trying to make use of citizen science to monitor ecosystems (e.g. Meyerhoff et al. 2003).

1.2 Our approach

We knew from our experience building Swoogle, a semantic web search engine (Ding et al. 2004), and TripleShop (Sachs et al. 2006), a SPARQL workbench, that ecological data represented in RDF or OWL can be discovered and integrated across datasets. Our previous work focused on transforming legacy data from large online databases or food web archives into RDF using automated scripts and several ontologies we developed (Parr et al. 2006).

The species-based ontology ETHAN (Evolutionary Trees and Natural History), underpins all of our subsequent work on organismal biology (Parr et al. submitted). ETHAN includes a subsumption hierarchy for the taxonomic or phylogenetic hierarchy of organismal names. It also includes subclasses grouping those organisms by their ecological, geographic, physiological, and physical characteristics. We use ETHAN as we create ontologies to support observations of species and various governmental lists of species of conservation or invasive species concern.

A logical next step was to work on a fast and convenient way to discover, integrate, store and analyze invasive species reports as they are generated. Because of the need to cast a wide net, it makes sense to use existing non-semantic resources, especially those where large numbers of users already contribute information which might be highly relevant to ecological research. Our goal is to enable users to quickly generate lightweight, OWL-based semantic web documents as supplements to their more traditional content on social websites, as described below.

1.3 The potential of social web sites

In recent years photosharing sites have become popular. Users of Flickr[1] upload thousands of pictures every day. Some of those pictures – namely those showing species of interest with some additional information like location, comments or captions with additional description - are potentially useful sources of information for ecologists.

Weblogs, or blogs are publicly viewable journals to which an individual or group of individuals posts frequent entries organized by date. Many bloggers use them to chronicle their daily lives, and for the subgenre of ecoblogs this includes what they observe about the natural world around them. In Ecoblogs users report interesting observations, post pictures of plants, animals, details on what kind of animals or plants they observed, how many they saw, etc. Our project maintains a blog called FieldMarking where we post such observations (http://ebiquity.umbc.edu/fieldmarking).

Blogs and photosharing sites like Flickr have become successful largely because of their support for online communities, encouraging free discussion and cross-linking among users and content. In April there were over 14 million geotagged photos on Flickr, and millions of blogs in the blogosphere. There are at least 40 blogs devoted exclusively to ecological topics (Bruno 2007); not all currently report observations, but many non-ecological blogs probably do this

The HTML format of blogs makes it hard to harvest, and analyze information, follow trends, and integrate with other resources. Any kind of analysis involves a human browsing through the blogs or using the crude tag searching capability of a site such as Technorati[2] By allowing ordinary users to add semantics to largely unstructured data we aim to make these sources useful to invasive species science. Just as many blog authors have become citizen journalists, providing major news sources with leads and on-the-ground reporting, they can provide surveillance for invasive species or emerging diseases.

2. Related Work

Semantic blogging is not a new idea (Cayzer, 2004; Karger and Quan 2004). Previous efforts, however, have focused more on semantically explicit metadata than on scientific data itself. Perhaps the closest effort, in spirit, to our own is Data Blogging (Reger 2006). In fact, our first version of Fieldmarking used the reger.com data blogging platform. Ultimately, however, it was not flexible enough for our needs, and lacked sufficient support for RDF.

There are several existing efforts to add semantics to web resources. Flickr users and bloggers can tag their photos or posts with any text string. Emerging from these collections of tags are “folksonomies,” or loose vocabularies that suit the organizational needs of individual users, or small communities of users (Mathes 2004). An advantage of social tagging is the ease with which they can be assigned as that users and applications need not consult a centralized standard. Disambiguating or relating tags, however, is a challenge because they are uncontrolled text strings. Flickr has a clustering algorithm that provides some semantic grouping functionality with its tags (Begelman et al. 2006).

In Flickr, users can also add specialized tags to images such as geotags, which encode geospatial coordinates in a standardized way. A number of other “machine tag” formats are now supported by Flickr. Many are designed to be automatically generated by special applications.

There are currently several approaches for adding structure and/or semantics to web pages such as blog posts. Here we will focus on structured blogging, microformats, and RDFa.

The goals of structured blogging (StructuredBlogging.org) are very similar to ours, namely to enable the harvesting and syndication of blog entries describing reviews, events, and other structured data. Templates can be designed and deployed using platform-specific plug-ins, such as for WordPress. Users fill out the template and the result is posted on the both as human-readable text and as xml. Transformations exist to map the xml into RDF. We considered creating a species observation template (called “microcontent” in the structured blogging parlance), but decided instead to work directly with RDF/OWL.

Microformats is an approach to embedding semantics directly in XHTML, making use of existing HTML attributes (Khare and Celik 2006) A “species” microformat is currently under discussion (Microformats.org). Currently, the semantics of microformats are fairly light, and do not include the ability to relate concepts to one another. This becomes a problem if communities are unable or uninterested in adhering to a single, simple standard.

In principal, however, microformat tags can be organized via an OWL ontology, and we expect to incorporate the eventual species microformat in our work. Recent work has also shown the benefits of tying folksobnomies to ontologies (Passant 2007), although this requires additional effort on the part of both administrators and users.

RDFa (Adida and Birbeck 2006) is an alternative to the usual XML RDF syntax. It introduces new XHTML attributes, and uses them to express RDF. We did consider using RDFa for Fieldmarking, but preferred a clean separation of the human and machine-readable text.

Our approach differs from these others in two respects. First, our tools are explicitly tied to OWL ontologies. This maximizes the expressiveness possible in the resulting datastream. Secondly, we aim for platform independent tools that maximize the ability of users to capture

[1] http://flickr.com

[2] http://technorati.com

structured data both on their own posts and on the posts of others, and to store that structured data anywhere.

3. Splickr

Splickr (http://spire.umbc.edu/splickr) is an application for querying the Flickr database of publicly available pictures. It provides a convenient way of retrieve and browse photographs using tags and location. Results can be stored in semantic web format (OWL) which can later be queried using Semantic Web tools.

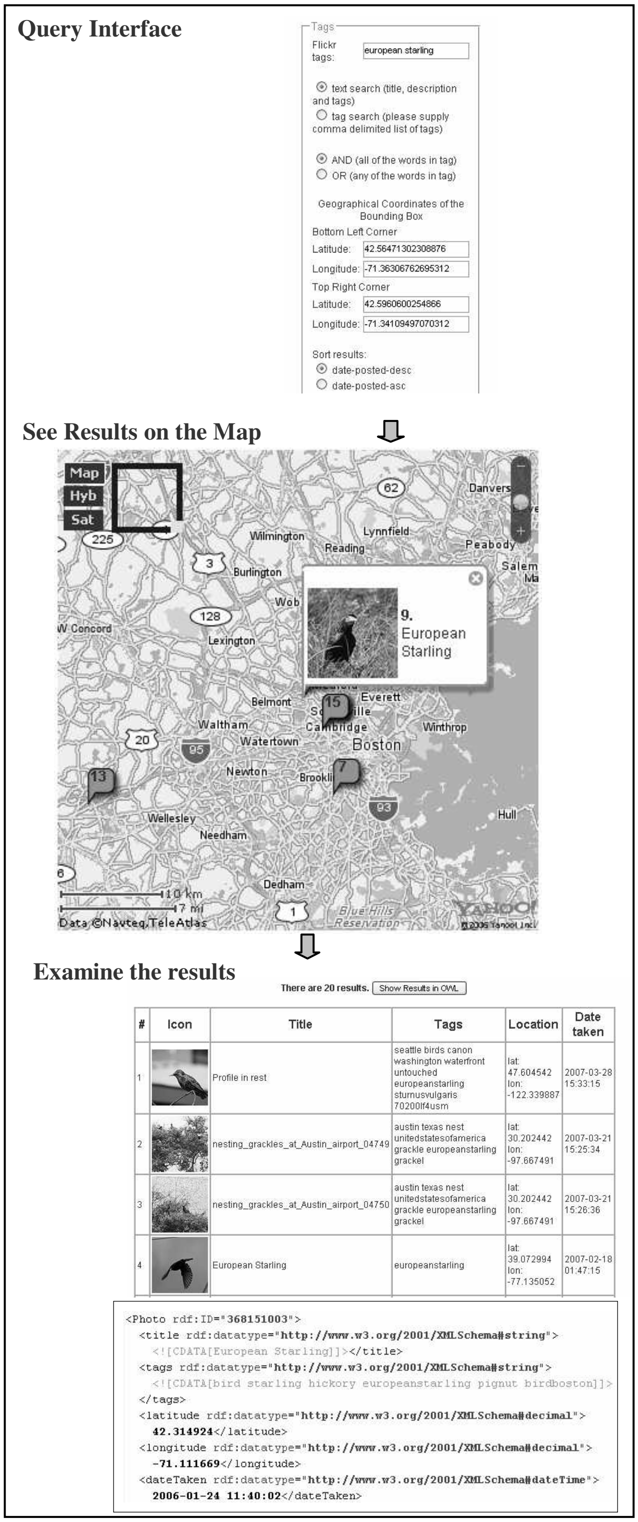
Figure 1 demonstrates finding a known invasive species, European starling, in Boston area.

The user enters the name of a species of interest, European starling, into the “Flickr tags” box. Selecting the text search option means that the search will look for the keywords anywhere on Flickr page – in the picture title, tags, comments, description etc. The user also specifies a sort preference is also specified. Splickr makes it convenient to define a geographical area of a search. Using the YahooMaps API (Yahoo Maps), we display a map on which a user can select a rectangular area of search, with the coordinates of the selected area automatically entered into corresponding boxes of Splickr search interface.

Results are displayed on the map as markers. By clicking on a marker, a user can see the icon of a picture and its title, clicking on an icon takes user to the photo’s full Flickr page. A summary of search results is also given HTML table below the map.

The processing and analyzing of the search results can be significantly enhanced by adding a semantic web context to the application. We developed a Splickr.owl ontology to express the search results in OWL. Figure 1 shows how one of the photos in the results is described using Splickr.owl. If these OWL-formatted results are saved in a public web directory, they can be indexed and retrieved by semantic web tools like Swoogle and TripleShop. They can thus be combined with other Semantic Web resources (e.g. SpireEcoConcepts ontology, ETHAN ontology suite etc) as shown in Figure 2.

The Splickr design and implementation are based on an AJAX technology which makes the application very interactive and user friendly. For example, the map interacts with the query interface so that geographical coordinates are entered automatically when user selects certain area and zoom level on the map. Once a query is submitted, an AJAX call to Flickr API is made and the Splickr page gets updated to display the results of the search. Thus, the entire session is a set of dynamic AJAX calls and updates so that the web page itself is never changed or reloaded, which helps to keep track of what a user has done before or is doing currently to provide the best user experience.



Tags

Flickr tags: european starling

● text search (title, description and tags)
○ tag search (please supply comma delimited list of tags)

● AND (all of the words in tag)
○ OR (any of the words in tag)

Geographical Coordinates of the Bounding Box

Bottom Left Corner
Latitude: 42.56471302308876
Longitude: -71.36306762695312

Top Right Corner
Latitude: 42.5960600254866
Longitude: -71.34109497070312

Sort results:
● date-posted-desc
○ date-posted-asc

There are 20 results.

#	Icon	Title	Tags	Location	Date taken
1		Profile in rest	seattle birds canon washington waterfront untouched europeanstarling sturnusvulgaris 70200lf4usm	lat: 47.604542 lon: -122.339887	2007-03-28 15:33:15
2		nesting_grackles_at_Austin_airport_04749	austin texas nest unitedstatesofamerica grackle europeanstarling grackel	lat: 30.202442 lon: -97.667491	2007-03-21 15:25:34
3		nesting_grackles_at_Austin_airport_04750	austin texas nest unitedstatesofamerica grackle europeanstarling grackel	lat: 30.202442 lon: -97.667491	2007-03-21 15:26:36
4		European Starling	europeanstarling	lat: 39.072994 lon: -77.135052	2007-02-18 01:47:15

```
<Photo rdf:ID="368151003">
  <title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    <![CDATA[European Starling]]></title>
  <tags rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    <![CDATA[bird starling hickory europeanstarling pignut birdboston]]>
  </tags>
  <latitude rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">
    42.314924</latitude>
  <longitude rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">
    -71.111669</longitude>
  <dateTaken rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">
    2006-01-24 11:40:02</dateTaken>
```

Figure 1. Splickr interface and usage. A user begins by entering her search term, and drawing a bounding box. Results are shown as markers on the map, and also in tabular form below. Each result can link to its OWL code.

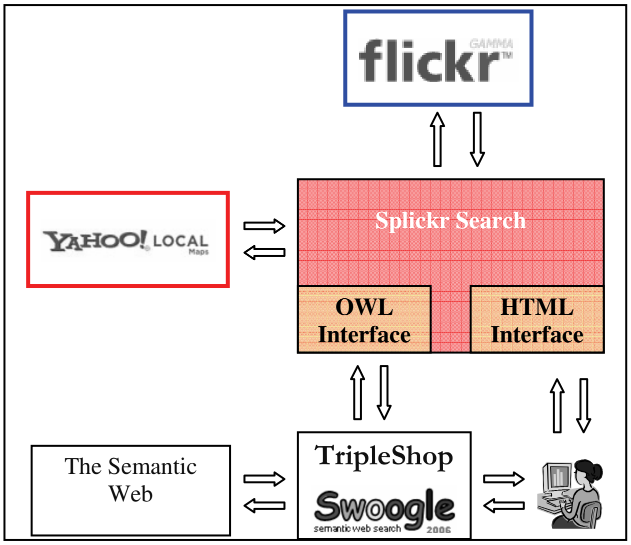


Figure 2 Splickr architecture. For our current purpose, the most relevant parts of the semantic web (lower left corner) are the ETHAN and SpireEcoConcepts ontologies and instance data.

4. Firefox plug-in

The SPire Observation Tool (SPOTter) Firefox extension (http://spire.umbc.edu/firefox) is another way to promote Semantic Web technology among ordinary users.

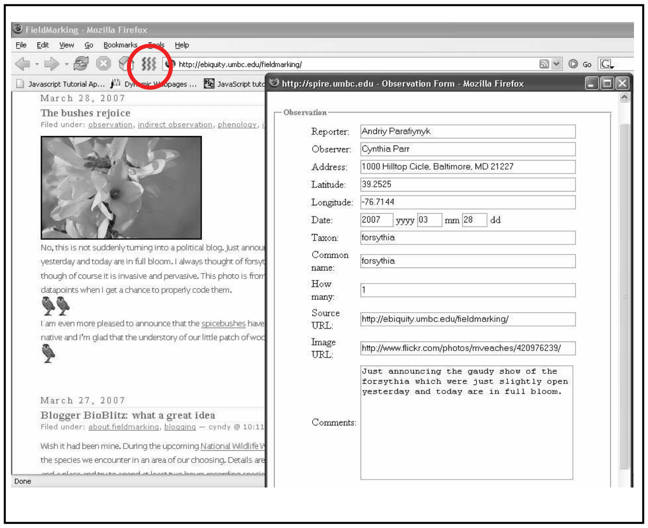
Most web users do not have the time, desire and necessary knowledge to write well-formed RDF documents. Yet the need for well-structured web data is already here. Having such RDF documents gives general and professional web users opportunity to use powerful semantic web tools (like Swoogle or TripleShop) to help in their daily activities or research. We seek to solve this problem by providing ordinary users with easy-to-install and easy-to-use tools for creating and using Semantic Web documents.

Semantic ecoblogging is a way to alleviate this problem and make blogs much more useful for scientific analysis. It involves creating semantic web content in addition to ordinary HTML text entries. Thus, the usual HTML content is in place for people who want to read the entire posts, look at the pictures etc, but there is also RDF data (which is better suited for aggregated analysis) for intelligent agents, applications or scientists who want to take advantage of automated processing.

SPOTter was created following the standard guidelines for building a Firefox extension, and installation is automatically performed by Firefox. Once installed, a SPOTter button appears on the Firefox tool bar and a new menu item is added to the main menu "Tools" section. This is all a user needs to know to use SPOTter.

Writing an observation on a blog, a user might want to save it in RDF as well. Clicking on the SPOTter button opens a simple form. Clicking on the button guides the user to the Spire server which directs the creation of the form and processes it when submitted.

The SPOTter form contains standard fields for reporting an observation. We developed an observation ontology (http://spire.umbc.edu/ontologies/Observation.owl) where all necessary OWL classes and properties are defined; organism names are from the ETHAN namespace. When the SPOTter form is submitted, the data are stored in a mySQL database (currently this service is provided for everyone for free) and user gets back an HTML link to the saved document. OWL documents (an OWL instance of the Observation class created based on the class definition in Observation.owl and the information provided by the user) are generated by a PHP script on demand. A user can see the submitted OWL observation by clicking on the link, and can copy and paste the link into his or her blog as a part of the blog entry. In the FieldMarking blog we use an icon of an owl to mark where OWL data is available. This link can later be followed and indexed by semantic search engines.


FieldMarking - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://ebiquity.umbc.edu/fieldmarking/
Go
Javascript Tutorial Ap...
Dynamic Webpages ...
JavaScript tuto
Done
http://spire.umbc.edu - Observation Form - Mozilla Firefox

March 28, 2007

The bushes rejoice

Filed under: observation, indirect observation, phenology,

No, this is not suddenly turning into a political blog. Just annou
yesterday and today are in full bloom. I always thought of forsyt
though of course it is invasive and pervasive. This photo is from
datapoints when I get a chance to properly code them.

I am even more pleased to announce that the spicebushes have
native and I'm glad that the understory of our little patch of woo

March 27, 2007

Blogger BioBlitz: what a great idea

Filed under: about fieldmarking, blogging — cyndy @ 10:11

Wish it had been mine. During the upcoming National Wildlife W
the species we encounter in an area of our choosing. Details are

Observation

Reporter:	Andriy Parafiynyk
Observer:	Cynthia Parr
Address:	1000 Hilltop Cicle, Baltimore, MD 21227
Latitude:	39.2525
Longitude:	-76.7144
Date:	2007 yyyy 03 mm 28 dd
Taxon:	forsythia
Common name:	forsythia
How many:	1
Source URL:	http://ebiquity.umbc.edu/fieldmarking/
Image URL:	http://www.flickr.com/photos/mveaches/420976239/
Comments:	Just announcing the gaudy show of the forsythia which were just slightly open yesterday and today are in full bloom.

Figure 3. SPOTter form. All fields are optional. Link to blog post is shown with owl icon. Button for launching is shown in red circle

Storing the data in a database and generating OWL on demand (as opposed to creating and storing static OWL documents) has several advantages:

1. It is easy to edit the SPOTter records, so OWL documents once created can be modified if needed. We provide an interface within the SPOTter Firefox extension to edit previous entries. With a PHP script calling mySQL database on the back end, previous records can be retrieved and easily updated.
2. The amount of disk space needed to store records is significantly smaller since only the values of OWL properties of an Observation instance are stored in the database, the rest of OWL (namespace declarations,

datatype declarations, properties etc) is added by the PHP template that generates OWL.

3. Storing the data in a mySQL database allows efficient querying of the data – it is fast and easy to find records matching SQL query constraints and deliver it to other applications. Using this functionality, we developed a SPOTter map (using YahooMaps API) which shows on a map all the SPOTter records matching certain search criteria (for example, observations reported within certain geographical area). Internally, the SPOTter map web page architecture is very similar to the Splickr web page architecture except that it makes calls to MySQL database instead of calls to the Flickr API. The SPOTter map web page is completely AJAX-based: the search panel, the Yahoo map and the database script interact with one another asynchronously using an AJAX approach; PHP calls to the database do not require reloading the page or map, nor does updating markers on the map. The map and results table updates are also asynchronous.

Currently, we have Swoogle crawling all observations made in two different blogs, FieldMarking (http://ebiquity.umbc.edu/fieldmarking) and Feathers of Hope (http://www.magpienest.org/feathersofhope/). Both of these blogs include primary observations as well as other kinds of posts. The intention is to foster SPOTter use on one's own blog and also to encode in RDF observations made by others. Using an icon or visible link indicates to others (and reminds oneself) where SPOTter code is available. This generic indication that a SPOTter record has been generated means that it can be added to virtually any web resource that allows comments with links. The ability to store the RDF record in any web-accessible directory similarly gives optimal cross-platform flexibility.

5. Future directions

These tools are still actively under development and will be made more usable as we partner with potential users. We plan to pilot test the tools during the upcoming Blogger's BioBlitz (Bruno 2007). We will also ask individual ecobloggers to try them out, and plan to work with students and staff of a local environmental education center. Ultimately, these tools should allow users of many backgrounds to encode and display their own data and that of others.

We plan a number of specific improvements. These include a preferences panel for designating user-specified default values and RDF-file locations, and to customize the link text a user wants to paste into their blog entry. Pre-filling out the form with data from an existing blog post and customization for common blog platforms may also be useful. We expect to add data to RSS feeds, and create a customizable map display for the SPOTter observations. Such a map should also access data generated by Splickr and could be embedded in a blog or other web page. A more advanced tool will allow users to specify different ontologies whose terms may be included in their SPOTter reports.

Splickr and SPOTter OWL documents are intended to be only a part of a bigger OWL knowledge base. Other OWL documents provide additional details on the species involved, for instance whether the newly reported species are invasive or not, which other species are potential prey and whether potential prey is on a list of endangered species etc. Ontologies will also allow intelligent expansion of search terms in Splickr.

We plan to compare our approach to generating semantic datastreams to that of other available technologies. We will consider ease-of-use, flexibility and customizability, potential for data integration and automation of retrieval.

Automating invasive species detection and monitoring using these tools will involve creating agents that watch these new streams of data, which will be indexed by Swoogle. Via intelligent SPARQL queries in TripleShop, the agents can identify observations of interest to scientists because they are known invasives spreading into a new area.

Although our case study domain is ecological data and invasive species in particular, the semantic web framework of these tools could be useful for many other domains where informal reporting may be important. These include emerging human or wildlife diseases, pollution sources, and reporting of suspicious criminal activity. Most or all of these applications would require integration with customizable trust filters.

In closing, this research is not intended to replace existing formal citizen science monitoring programs. Rather, these tools can augment existing programs and provide a way to harness observations that are made outside those protocols. Our vision is that these tools may identify important trends that will be worth more detailed and tightly controlled studies.

6. Acknowledgements

This research was supported by NSF ITR 0326460 and matching funds received from USGS Nat. Bio. Information Infrastructure. We thank Lushan Han and David Wang for helpful discussion. Allan Hollander was even brave enough to try out SPOTter on his own blog.

7. References

Adida, B. and Birbeck, M, RDF/A Primer 1.0 -- Embedding RDF in XHTML, W3C Working Draft 10, W3C, March 2006, available as http://www.w3.org/TR/xhtml-rdfa-primer/

Begelman, G. et al. 2006. Automated Tag Clustering: Improving search and exploration in the tag space. WWW'06

Bruno, J. 2007. The Voltage Gate http://scienceblogs.com/voltagegate/2007/04/blogger_bioblitz_updates_ii.php

Cayzer, S. 2004. Semantic blogging and decentralized knowledge management. Communications of the ACM 47(12),:47--52.ACM Press, New York, NY, USA

Ding, L.; Finin, T.; Joshi, A.; Pan, R.; Cost, R.S.; Peng, Y.; Reddivari, P.; Doshi, V.; and Sachs, J., 2004. Swoogle: a search and metadata engine for the semantic web. Proceedings of the thirteenth ACM international conference on Information and knowledge management p. 652-659. ACM Press, New York. Swoogle available at: http://swoogle.umbc.edu

Karger, D.R. and Quan, D. What would it mean to blog on the semantic web. Proceedings of the Third International Semantic Web Conference (ISWC2004), Hiroshima, Japan.

Khare, R. and Celik, T. 2006. Microformats: a Pragmatic Path to the Semantic Web. Proceedings of the 15th International Conference on the World Wide Web (WWW2006). ACM Press, New York, NY, USA.

Mathes, A. 2004. Folksonomies-Cooperative Classification and Communication Through Shared Metadata. Available at http://blog.namics.com/archives/2005/Folksonomies_Cooperative_Classification.pdf

Mehrhoff, L. J.; Silander, Jr., J. A ; Leicht, S. A.; Mosher, E. S.; and Tabak, N. M. 2003. IPANE: Invasive Plant Atlas of New England. Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, CT, USA. Sighting form available at http://www.ipane.org (IPANE). Sighting form: http://nbii-nin.ciesin.columbia.edu/ipane/earlydetection/sightings.jsp

Microformats.org. Available at http://microformats.org/ wiki/species

Parr et al submitted., ETHAN: the Evolutionary Trees and Natural History Ontology. Ecological informatics.

Parr, C. S.; Parafiynyk, A; Sachs, J.; Pan, R.; Han, L.; Ding, L.; Finin, T.; Wang, T.D.; and Hollander. A.. 2006. Using the semantic web to integrate ecoinformatics resources. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06), July 2006; 1949-1950. TripleShop available at: http://sparql.cs.umbc.edu/tripleshop2/

Passant, A, 2007. Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs: Theoretical background and corporate use-case. International Conference on Weblogs and Social Media. Boulder, CO March 2007.

Pimental, D.; Lach, L.; Zuniga, R.; Morrison, D. 2000. Environmental and economic costs associated with non-indigenous species in the United States. Bioscience 50:53-65.

Reger, J. 2006. http://reger.com/about/what-is-datablogging.log

Sachs, J, Cynthia Parr, Andriy Parafiynyk, Rong Pan, Lushan Han, Li Ding, Tim Finin, Allan Hollander, Taowei Wang. Using the Semantic Web to Support Ecoinformatics. Proceedings of the AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition, October 13, 2006.

StructuredBlogging.org. http://structuredblogging.org

Yahoo! Maps. http://developer.yahoo.com/maps/