Bioblitz Data on the Semantic Web

Joel Sachs
C.S.E.E.
niversity of Maryland Baltimore County
Baltimore, MD, 21250 USA
jsachs@umbc.edu

Lushan Han
C.S.E.E.
University of Maryland Baltimore County
Baltimore, MD, 21250 USA
lushan1@umbc.edu

David Wang
Computer Science
University of Maryland
College Park, MD, 20742 USA
Tw7@cs.umd.edu

Cynthia Simms Parr
Encyclopedia of Life
Smithsonian Institution
Washington D.C., 20013 USA
parr@si.edu

Abstract

Last summer, we represented the 1200 species observations of the First Annual Blogger Bioblitz in RDF. This allowed us to easily integrate the data with existing RDF data on natural history, taxonomy, food webs, conservation status, and invasive status. Using our Swoogle/Tripleshop approach to dataset construction, we were able to respond to a variety of ad-hoc queries

Our efforts last year were external to the running of the bioblitz. For this year's blogger bioblitz (late August), we have taken responsibility for data processing, and will encourage participants to make use of two tools we have developed that ease the process of user-generated RDF – the Spotter Firefox plug-in, and RDF123. We will encourage ISWC participants to photo-blog Karlsruhe wildlife during the conference, and to use Spotter to generate RDF of their posts. We will also attempt a full bioblitz of a suitable area near the Conference Centre. Our demo will allow users to browse the dataspace resulting from all observations, together with background data, and to issue SPARQL queries over the data.

1. Background

1.1 Semantic Eco-blogging and Spotter

Eco-blogs are popular amongst both amateur nature lovers and working biologists. Subject matter varies, but entries typically include date, location, observed taxa, and description of behavior. These observations can be an important part of the ecological record, especially in domains (such as invasive species science) where amateur reporting plays an important role, and in the study of environmental response to climate change.

To enable our goal of a human sensor net, we developed Spotter [1], a Firefox extension that enables the easy creation of RDF data by citizen scientists. Spotter is not tied to a particular blogging platform, and can be used both to add semantic markup to one's own blog posts, and to annotate posts or images on other websites, such as Flickr.

Once RDF is generated, we apply much of the machinery we have developed as part of the SPIRE project - Swoogle, our Semantic Web search engine; Tripleshop, our distributed dataset constructor; and ETHAN, our evolutionary trees and natural history ontology. We are then able to issue queries like:

What was the northernmost spotting of the Emerald Ash Borer last year?; Show all sightings of invasive plants in California; etc.

1.2 The blogger bioblitz

A bioblitz is a 24 hour inventory of the biodiversity of a particular area, such as a park. The primary purpose of bioblitzes has traditionally been scientific outreach. However, as they have gained popularity in recent years, they are increasingly being looked at as important sources of biodiversity data. The U.S. National Biological Information Infrastructure, for example, has expressed an interest in hosting U.S. bioblitz data.

Last year, eco-blogger Jeremy Bruno organized the first blogger bioblitz. Eco-bloggers from around the world chose a day during a specified week, and did an inventory of the species in an area near them. Over 30 bloggers participated, with 17 submitting datasheets to Jeremy for processing. We expressed the 1200 reported observations in RDF, and, using our Swoogle/Tripleshop approach, were able to respond to a number of ad/hoc queries. For example, when we issued the query

Show all observations of species that are classified as being of concern (i.e. invasive, threatened, etc.),

we got back 47 records.

More generally, our experience illustrated how scientists can share data by annotating it with RDF, publishing it via plug-ins to popular software, and making it accessible via new tools and Web mashups.

1.3 RDF123

We developed RDF123 [2] as a highly flexible open-source tool for transforming spreadsheet data to RDF. Our work was, and continues to be, motivated by the fact that spreadsheets are easy to understand and use, offer intuitive interfaces, and have representational power adequate for most purposes. Moreover, online spreadsheets are increasingly popular and have the potential to boost the growth of the Semantic Web by providing well-formed and publicly shared data sources that can be directly maintained by users and automatically translated into RDF.

In our approach, we borrow the idea from GRDDL [3] of placing a link in an online spreadsheet referencing the map file, which is itself an RDF document, specifying the desired translation. When an agent comes to the spreadsheet, it follows the link, reads the map file, applies it to the spreadsheet and thus generates RDF data. Moreover, RDF123's Web service also allows users to apply map files to other users' online spreadsheets and generate their customized RDF data.

We consider the introduction of the RDF123 web service to be a significant step in the assimilation of spreadsheets into the semantic web. This service takes as input the URI of a spreadsheet and the URI of a map, and gives RDF as output. Thus RDF is generated on the fly, obviating the need for data dumps, and the resultant stale data. Scientists can, therefore, maintain their spreadsheets as usual, with all updates becoming an immediate part of the semantic web. This fulfills one of the original goals for science on the semantic web, namely that semantic markup should fall out of the everyday work processes of scientists.

1.4 SPIRE data

We expect that many opportunities for integrating observations with the ever-growing web of data will present themselves. As a starting point, the following SPIRE data is directly relevant:

ELVIS. ELVIS is motivated by the belief that food web structure plays a role in the success or failure of potential species invasions. Because very few ecosystems have been the subject of empirical food web studies, response teams are typically unable to get quick answers to questions like "what are likely prey and predator species of the invader in the new environment?" Our ELVIS tools [4] seek to fill this gap by using trophic links gleaned from over 250 food web studies to predict food web structure in unstudied ecosystems. ELVIS expresses all data in OWL via our ecological and evolutionary ontologies.

ETHAN. The species-based ontology ETHAN (Evolutionary Trees and Natural History), underpins all of our work on organismal biology. ETHAN includes a subsumption hierarchy for the taxonomic (or, in some cases, phylogenetic) hierarchy of organismal names. It also includes subclasses grouping those organisms by their ecological, geographic, physiological, and physical characteristics. Quantitative measures such as body mass and lifespan are also represented. We continue to use ETHAN as we create ontologies to support observations of species and various governmental lists of species of conservation or invasive species concern.

Conservation Information in OWL. An illustration of the flexibility of our approach is the way we handle conservation designations such as invasive, endangered, etc. Typically, a species if designated invasive through inclusion in a list defined by multi-lateral agreement, or by legislation. For each such list we want to consider, we create a corresponding OWL class. Species on a particular list are then asserted to belong to the corresponding class. This can be seen by browsing our Invasives ontology[1] and the classes it defines[2].

Using this approach, we were, for example, easily able to introduce the issue of invasiveness into all our previous SPIRE work.

[1] http://spire.umbc.edu/ontologies/InvasivesOntology.owl
[2] http://spire.umbc.edu/ontologies/lists/

2. The 2008 blogger bioblitz

2.1 Data Generation

Processing the data was a headache for last year's organizer, and he is keen on outsourcing data analysis to us this year. Bloggers will be able to contribute data in a number of ways: they can use Spotter to generate an RDF record for each taxon they observe; they can maintain an on-line spreadsheet, which will be automatically converted (upon http request) to rdf ; or they can email us their spreadsheet, which we will convert with RDF123. This will be the first large scale RDF123 experiment, and we are excited about the prospects not only for collecting a large amount of user-generated RDF, but also for demonstrating the opportunities for merging spreadsheets that RDF123 provides.

2.2 Querying Capabilities

We will provide a variety of views on the data, stressing a linked data approach to enable data browsing. A user will, at a minimum, be able to
i. browse bioblitz data together with other linked natural history and ecology data;
ii. view standard bioblitz summary statistics, e.g. counts by taxon, map views, etc.
iii. issue queries of the form "Show all observation of species from <*some list*>?", where *some list* could be *Threatened species*; *Invasive plants of California*; *etc*.

We are also hoping to generate views of the data based on phenology and other potential climate change indicators.

As well, we will experiment with a spreadsheet-based query-by-example approach. Users will partially fill out a spreadsheet, which will then define a set of SPARQL queries to fill in the remaining cells.

2.3 Eco-blogging Karlsruhe

We will encourage ISWC participants to post photographs of Karlsruhe wildlife either on their own blogs, the conference blog, or our Fieldmarking site, and to use Spotter to generate RDF of their posts. We will also attempt a full bioblitz of a suitable area of Karlsruhe, possibly the Statdgarten, which is immediately South of the Congress Center.

3. Acknowledgements

This research was supported by NSF ITR 0326460

References

[1] Andriy Parafiynyk et al., "Adding Semantics to Social Websites for Citizen Science", Proceedings of the Workshop on Semantic e-Science (AAAI 07), June 2007.

[2] Lushan Han et al., "RDF123: a mechanism to transform spreadsheets to RDF", Proceedings of the 7th International Semantic Web Conference (ISWC 2008 – to appear)

[3] D. Hazael-Massieux and D. Connolly, "Gleaning Resource Descriptions from Dialects of Languages" (GRDDL), Coordination Group Note NOTE-grddl-20040413, World Wide Web Consortium, April, 2004.

[4] Joel Sachs et al., "Using the Semantic Web to Support Ecoinformatics", Proceedings of the AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition , October 2006.