

# SemantEco Annotator

Patrice Seyed<sup>1,2</sup>, Katherine Chastain<sup>2</sup>, Brendan Ashby<sup>2</sup>, Yue Liu<sup>2</sup>,  
Timothy Lebo<sup>2</sup>, Evan Patton<sup>2</sup>, and Deborah McGuinness<sup>2</sup>

<sup>1</sup>DataONE, University of New Mexico, Albuquerque, NM 87131

<sup>2</sup>Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA  
{seyeda2, chastk, ashbyb, liuy18, lebot, pattoe, dlm}@rpi.edu

**Abstract.** Generating useful RDF linked data is not a straightforward process for scientists using today's tools. In this paper we introduce the SemantEco Annotator, a semantic web application that leverages community-based vocabularies and ontologies during the translation process itself to ease the process of drawing out implicit relationships in tabular data so that they may be immediately available for use within the LOD cloud. Our goal for the SemantEco Annotator is to make advanced RDF translation techniques available to the layperson.

## 1 Introduction

Scientists generating datasets of tabular data make choices about how they record that data. These decisions are informed by many factors, including how scientists understand and analyze the data. Often, information that can be gleaned from a table alone is limited by the initial input strategy, including what is appropriate for the table headers (i.e., attributes) and cell values, and the conventions for structure and terminology. There may be implicit information apparent only to the creator, leading to challenges for uniformly interpreting data generated by different researchers. Linked data formats based on web standards such as RDF provide an avenue for addressing this problem, allowing the data to be more explicitly represented in statements uniquely typing entities and relationships described in data using common vocabularies. Unfortunately, the process for translating source data into linked data has its own barriers, as software tools allowing scientists that are inexperienced with linked data to perform translations are lacking.

In our demonstration we introduce SemantEco Annotator, a semantic web application for translating tabular data into RDF for immediate use in the LOD cloud. The annotator serves as a frontend to the *csv2rdf4lod* [1] conversion tool that uses an RDF metadata vocabulary to specify a variety of powerful techniques for translating data from CSV to RDF. We see the SemantEco Annotator as an application that makes translation more accessible to the layperson, and which can act as a module in a larger semantic workbench. We created our initial implementation for our SemantEco environment, which uses linked data to present water quality data in a visual and interactive manner [2]. In the next section we describe other RDF translation tools, following which we highlight the primary translation techniques the

annotator currently enables, and finally we provide some insight into how the application was developed, and our future work.

## 2 Related Work

There are existing tools both for converting tabular data to RDF and for semantic annotation. RDF Refine<sup>1</sup> is an extension to OpenRefine where conversion to linked data requires that a user 1) define a “skeleton” for the final RDF as a tree structure, 2) maps table columns to nodes of the tree. Our tool combines these steps, providing a simpler interface by keeping all of the interactions for RDF mapping situated with the original data table. The users also retain the flexibility to define or change the schema as they go along. RDF Refine also allows certain types of conversions through “transposition” that requires the user to change the structure of the original tabular data in order to perform the conversion. Our approach enables similar modifications while maintaining the original structure of the CSV file.

Anzo Express<sup>2</sup> by Cambridge Semantics enables semantic annotation of Excel spreadsheets, and also provides various other functions, including an OWL ontology editor. Our tool focuses on re-use of existing ontologies and community-standard vocabularies. In contrast to both RDF Refine and Anzo Express, *csv2rdf4lod* has the added functionality of generating provenance statements using popular vocabularies<sup>3</sup> that enable tracking the original dataset through each of the stages from retrieval, through conversion, and its publication as dump files or into a SPARQL endpoint.

## 3 Introducing the SemantEco Annotator

To illustrate the primary translation techniques available through the annotator, we consider data from the Darrin Freshwater Institute that tracks the quality of lake water in New York State’s Adirondack region. The techniques are 1) row and cell-based translation; 2) implicit and explicit bundling; and 3) leveraging OWL ontology classes, properties, and individuals to formulate new RDF statements. Each of these techniques is demonstrated and discussed in detail in our video linked at <https://github.com/apseyed/SemantEcoAnnotator/wiki>.

### 3.1 Row and Cell-Based Translation

A simple RDF translation of a table generates a triple for each non-empty cell value, where the subject, predicate, and object URIs are formed from the row, column, and cell value, respectively. The subject URI is either generated from a convention or

---

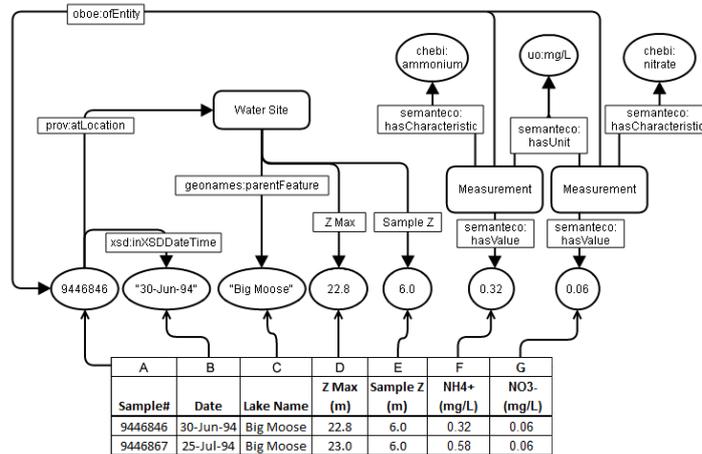
<sup>1</sup> <http://refine.deri.ie/>

<sup>2</sup> <http://www.cambridgesemantics.com/products/anzo-express>

<sup>3</sup> <http://prefix.cc/void,dcterms PROV>

constructed from a column's cell value. This default technique as *row-based translation*, since every triple generated from a row shares the same subject.

In more complex tabular arrangements, each row may instead contain many subjects that are described across multiple columns, and are related direct or indirectly. The technique for this pattern is *cell-based translation*. Figure 1 illustrates a scenario where it is useful for columns F and G, which contain measurements of NH<sub>4</sub><sup>+</sup> and NO<sub>3</sub>, respectively, for a water sample. Direct or indirect features of the measurements are contained in other cells in the same row, and relate back to the sample. Columns are designated for cell-based translation via a context menu. Translating the indirect relationships requires *bundling*, which we discuss in the next section.



**Fig. 1.** An excerpt of water quality data from Darrin Freshwater Institute including information about the sample (Cols. A-E) and two of its measurements (Cols. F,G). Directed arcs above the table illustrate relationships between what is described in columns, which our annotator helps a user describe and ultimately generates for the user via translation into RDF.

### 3.2 Bundling

Sometimes a column represents a description of an entity identifiable in another column. The Date (col. B) describes when the Sample (col. A) was taken; the user can capture this relationship by creating an *explicit bundle*. In contrast, columns may describe an unspecified, or implicit set of entity not captured by another. For the water sample data, the Lake Name (col. C), the depth to the bottom of the lake (col. D, Z Max), and the depth of sample collection (col. E, Sample Z) together represent a sample location. The user can represent the implicit entity by bundling the columns that represent it together, in an implicit bundle. Bundled columns are visually “pushed down” into a new row in the header to show they are now grouped, and subordinate aspects of the entity they are bundled into. After the bundling selection is complete, the user has a choice of one of the existing column identifiers for explicit from a drop-down list, or declaring it an implicit bundle.

For the water measurement columns (*F*, *G*) designated for cell-based conversion, there is vital unit information in the header. We provide a *subject annotation* feature to directly assert out-of-band triples about the measurement units. Annotations are created with the context menu and accept dragged properties and classes to generate further triples. We illustrate these steps in our demonstration video.

### 3.3 Enhancing RDF Translation with OWL Ontologies

In connection with the aforementioned translation techniques, we provide mechanisms for using OWL ontologies. The right side of the interface includes a catalog of ontologies to choose from, which in turn populates trees with choices of the ontologies' classes, properties, and datatypes, for adding semantics to the resultant RDF data. If an object, data, or annotation property node is dragged from the tree interface into a column, the triples generated for that column will use the given predicate. Dragging a class node into the column will type the object of the triple as an instance of the given class. We also employ restrictions to selections based on previous choices for a column. This is highlighted in our demonstration video, through use of various scientific-observation ontologies such as OBO-E<sup>4</sup>.

## 4 Discussion and Future Work

Once enhancements are committed RDF files for the enhancement parameters and the RDF produced by *csv2rdf4lod* are downloadable for use as LOD. In the future we will leverage logical restrictions from ontologies to guide the user to constructs that are most appropriate for subsequent enhancements. We also plan to extend the ontology menu to load any ontology listed in catalogs such as BioPortal<sup>3</sup>, as well as via URI. While the annotator enables translation of data into RDF using existing ontologies and vocabularies, other components can enable mappings of terms within data packages to ontology concepts. This environment would essentially enable generation of semantically enriched data as linked data to enable better search capabilities.

## References

1. Lebo, T., Erickson, J.S., Ding, L., Graves, A., Williams, G.T., DiFranzo, D., Li, X., Michaelis, J., Zheng, J., Flores, J., Shangquan, Z., McGuinness, D.L., and Hendler, J. Producing and Using Linked Open Government Data in the TWC LOGD Portal. In *Linking Government Data*, pages 51–72. New York, 2011. [10.1007/978-1-4614-1767-5\\_3](https://doi.org/10.1007/978-1-4614-1767-5_3)
2. Seyed, A. P., Lebo, T., Patton E., McCusker J., and McGuinness D. L. SemantEco: A Next-Generation Web Observatory. 1<sup>st</sup> International Web Observatory Workshop. 22<sup>nd</sup> International World Wide Web Conference, Rio de Janeiro, May 13-17, 2013.

---

<sup>3</sup><https://semttools.ecoinformatics.org/oboe>

<sup>4</sup><http://bioportal.bioontology.org/ontologies>