Metabolite Annotations and Semantic Web Integration with gLIMS

Edward Pharr
Computer Science Department
College of Charleston
66 George Street
Charleston, South Carolina 29424
Email: edwardpharr@gmail.com

Aspen Olmsted
Computer Science Department
College of Charleston
66 George Street
Charleston, South Carolina 29424
Email: olmsteda@cofc.edu

Paul E. Anderson
Computer Science Department
College of Charleston
66 George Street
Charleston, South Carolina 29424
Email: andersonp@cs.cofc.edu

Abstract—gLIMS is a laboratory information management system and community-driven spectroscopic annotation system for the field of metabolomics, which is the study of the products left behind by the process of metabolism. gLIMS provides services and interfaces for exposing nuclear magnetic resonance (NMR) spectral annotations to the Semantic Web. The ability to plot and annotate spectra enables the metabolomics community to improve analysis and subsequent scientific discovery. gLIMS is flexible, efficient, available, and easy to integrate with scientific workflow software by providing a representational state transfer (RESTful) web service interface that allows researchers to view, interpret, and then annotate significant metabolites on NMR spectroscopic data. gLIMS exposes spectra and metabolomics annotation metadata and provenance to the Semantic Web through the Resource Description Framework (RDF) which allows the system to integrate with Linked Data and be queried via SPARQL. This system is implemented using Google's platform as a service (PAAS) technologies and Google Drive to produce a robust, scalable, and extensible lab information management system running entirely in-browser which facilitates collaboration on a worldwide scale.

I. INTRODUCTION

The relatively young field of metabolomics is the measurement of metabolite concentrations and changes in various biological systems [2]. Metabolomics assesses the end product of cellular functions and reflects the cellular metabolic status. Nuclear magnetic resonance (NMR) spectroscopy effectively identifies variations in biological states and is not invasive or destructive, and it requires little sample preparation. Metabolomics experiments require substantial computational and statistical support, and researchers require significant infrastructure for storage, visualization, and peer-review of online repositories of this highly multidimensional data.

Metabolomics requires intensive data analysis for interpretation of experimental results, quantifying and annotating hundreds of metabolite levels for each sample analyzed [3]. NMR metabolic spectroscopic data are analyzed in a five step process: (1) post-instrumental processing of spectroscopic data, (2) quantification of spectral signals, (3) normalization and scaling, (4) multivariate statistical modeling, and (5) metabolite identification. The final step, metabolite identification, is a difficult process, but determining the elements present in a spectrum is the application in which gLIMS is particularly useful.

Metabolite identification is influenced by the peak alignment quality and breadth of the training database. The assembly of a comprehensive database of spectral annotations is complicated by dependence on sample composition, field strength, ionic strength, pH, and other experimental conditions. The lack of such a database significantly hinders the development of automatic annotation algorithms; however, such a database does exist in form of individual annotation libraries by domain experts of the international community of NMR scientists. gLIMS combines a data storage and sharing mechanism with software to perform identification and annotation. With gLIMS, scientists can upload their spectroscopic data and use the community curated database for metabolite identification. gLIMS connects open standards, extensibility, and shared data in a community driven laboratory information management system. The availability and usefulness of the data is furthered by our implementation of a semantic web interface with RDF graphs and SPARQL queries. The semantic web interface, which will include faceted search, will allow scientists to leverage inference without data science expertise. The target and impact of this application are broad, but our initial focus is nuclear magnetic resonance (NMR) metabolomics. In particular, toxicology datasets are available as a preliminary demonstration of the usefulness of our system[1].

II. SEMANTIC WEB AND METABOLITE ANNOTATIONS

The features of gLIMS were designed to support the NMR-based metabolomics community, specifically providing cyber-infrastructure, visualization, and annotation. Multiple datasets can be graphed simultaneously to contrast experimental results. Plots of spectroscopic data can be annotated and the resulting provenance stored and added to the repository providing valuable labels for machine learning algorithms to automatically identify metabolites.

When gLIMS parses a file stored on Google Drive, individual samples are written and stored in appropriate collections (tagged) based on the provenance associated with each sample. These provenance labels can be browsed, updated, and downloaded using the Drive user interface. The provenance hierarchy associated with each collection of data is preserved.
when data is viewed and graphed with gLIMS. RDF provides an alternative representation of the provenance expressed by our Drive collections and our metabolite annotations exposes labeled spectroscopic data to the Semantic Web. Collections, which function as tags, give context to the data, and dynamically generated RDF graphs expand that context further. This provides increased inference and querying capabilities.

Expressing our provenance as an RDF graph will allow the user to execute SPARQL queries on the community derived annotation and spectroscopic databases. Provenance associated with spectroscopic data are represented as subject, predicate, object triples which are both efficient and generalizable. RDF graphs are decentralized, distributed, and can be merged with other sources. Users can issue queries across multiple data sources and be returned one combined result set, allowing them to extend beyond the data they have stored on Drive to data hosted by others in the metabolomics community. For example, two species of coral, porites lobata and seriatopora hystrix, are linked through the class anthozoa. A partial schema and a sample query with results that filter samples collected during the day and over a certain weight are shown below.

(i) Data

```xml
<https://docs.google.com/file/d/0B7Jf...>
<https://docs.google.com/file/d/0B7Jf...>
drive:Proites_lobata rdf:subClassOf Anthozoa
drive:Seriatopora_hystrix rdf:subClassOf Anthozoa
<https://docs.google.com/file/d/0B7Jf...>
drive:extract_weight 0.0053;
    drive:sampling_time_point "day";
    drive:species Porites_lobata;
<https://docs.google.com/file/d/0B7Jf...>
drive:weight 0.0051;
    drive:sampling_time_point "day";
    drive:species Seriatopora_hystrix;
```

(ii) SPARQL Query

```
SELECT ?x WHERE {
  ?x drive:sampling_time_point "day".
  FILTER (?weight > 0.0050)
}
```

(iii) Query Result

```
<table>
<thead>
<tr>
<th>desc.</th>
<th>weight</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taiwan</td>
<td>0.0053</td>
<td>Porites lobata</td>
</tr>
<tr>
<td>Taiwan</td>
<td>0.0051</td>
<td>Seriatopora hystrix</td>
</tr>
</tbody>
</table>
```

The community generated annotation database can be fed to machine learning algorithms, providing a first step in an automated process to identify the elements present in a sample. The ability to peer-review existing annotations and compare annotated empirical spectroscopic data is critically important to metabolomics analysis. Scientists will be able to confirm or question the marked spectra. With SPARQL, scientists will be able to issue queries for instances of a metabolites and annotations that meet the particular criteria in which they are interested. The review process will improve the overall quality of the annotations and the effectiveness of algorithms applied to the data. Correctly labeled data will be useful to those trying to characterize metabolites under a diverse set of experimental conditions.

REFERENCES