Aggregation of Bioinformatics Data Using Semantic Web Technology

Susie Stephens[1*], David LaVigna[2], Mike DiLascio[2], Joanne Luciano[3]

[1] Oracle, 10 Van de Graaff Drive, Burlington, MA 01803, USA
[2] Siderean Software, Inc., 390 North Sepulvada Boulevard, Suite 2070, El Segundo, CA 90245-4475, USA
[3]Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Abstract
The integration of disparate biomedical data continues to be a challenge for drug discovery efforts. Semantic Web technologies provide the capability to more easily aggregate data and thus can be utilized to improve the efficiency of drug discovery. We describe an implementation of a Semantic Web infrastructure that utilizes the scalable Oracle RDF Data Model as the repository and Seamark Navigator for browsing and searching the data. The paper presents a use case that identifies gene biomarkers of interest and uses the Semantic Web infrastructure to annotate the data.

Keywords: Data Aggregation; Bioinformatics; Faceted Browsing; Oracle RDF Data Model; Semantic Web.

1. Introduction

To make well-informed decisions, biomedical researchers need to be able to easily access all relevant data. To achieve this goal, data about genes, proteins, pathways, diseases, and chemical compounds must be effectively integrated and readily available to the researcher.

The life sciences has a rich history of making data available on the Web. Early on, scientific researchers recognized the benefits of sharing their data and made it available to other researchers for the benefit of the greater good. However, because many of these data repositories were developed in relative isolation, it has resulted in a heterogeneous compute environment that makes it challenging for scientists to find and properly utilize all known information about an entity of interest. In this environment, scientists must jump from Web site to Web site following a path of interconnecting identifiers in order to find all necessary data. This process is difficult to automate because many Web sites do not have programmatic interfaces to their data and it is difficult to capture the scientific thought process.

Many articles have been written that highlight the challenges of data integration within the biomedical domain and illustrate the difficulty involved in integrating the many publicly available bioinformatics data sources with in-house ones. The challenges stem from the data sources having different identifier schemes, inconsistent terminology, multiple data formats, and regular changes to the underlying data models.

In hopes to lessen the burden of data integration and sharing, life sciences researchers and organizations have recently begun to explore Semantic Web technologies [1]. The benefits promised by the Semantic Web include aggregation of heterogeneous data using explicit semantics, simplified annotation and sharing of findings, the expression of rich and well-defined models for data aggregation and search, easier reuse of data in unanticipated ways, and the application of logic to infer additional insights [16].

The two main Semantic Web standards are Resource Description Framework (RDF) [14] and Web Ontology Language (OWL) [15]. RDF represents data using subject-predicate-object triples (also known as ‘statements’). This triple representation connects data in a flexible piece-by-piece and link-by-link fashion that forms a directed labeled graph. The components of each RDF statement can be identified using Uniform Resource Identifers (URIs). Alternatively, they can be referenced via links to RDF Schemas (RDFS), OWL ontologies, or to other (non-schema) RDF documents.

Organizations in the life sciences are currently using RDF for drug target assessment [http://www.olsug.org/wiki/images/d/df/AWL.pdf], and the aggregation of genomic data [2]. In addition, Semantic Web technologies are being used to develop well-defined and rich biomedical ontologies to assist with data integration and search [6, 7, 10]; the integration of rules to specify and implement bioinformatics workflows [9]; and the automation of discovery and composition of bioinformatics Web Services [21].

This paper provides an overview of two specific Semantic Web technologies, namely the Oracle RDF Data Model and Seamark Navigator from Siderean Software, and then provides an example as to how these technologies can be utilized for drug discovery.

[*] Corresponding author: tel. +1 781 744 0372; fax. +1 781 238 9857; email. susie.stephens@oracle.com

2.0 Software Components

2.1 Oracle RDF Data Model

In Oracle Database 10*g* Release 2, support is provided for RDF and RDFS. The implementation is based on the object relational capabilities of the database. All RDF triples are stored in the system as entries in tables, but the user interacts with the triples at an object level. Functionality is provided to enable users to link from the RDF Data Model to the multimedia capabilities within the database, thereby allowing images and textual documents to form a component of the graph.

The RDF Data Model has proven scalable due to the ability to reuse subject and object components of triples, and by allowing data to be partitioned into distinct models. Database features such as indexing, memory management, and parallelization can be used with the RDF functionality. In addition, the Oracle Real Application Clusters capability can be used with the RDF Data Model, enabling users to run the database instance over several nodes in a compute cluster. The scalability of the implementation has been analyzed using UniProt [3].

SQL has been extended to allow users to search for an arbitrary pattern within the RDF data. The implementation supports inferencing based on RDF, RDFS and user-defined rules [20]. The graph query capability is based upon the RDF query requirements as identified by the W3C Data Access Working Group [17].

2.2 Seamark Navigator

Seamark Navigator from Siderean Software is a Web application that allows users to browse, search and organize RDF data. Seamark can be used to generate RDF metadata from a number of data types, including tab-delimited files, relational data, XML, and RSS. Once the RDF metadata has been created, Seamark can be used to organize the disparate data into a graph.

Seamark provides faceted navigation capabilities to guide users to relevant content. A facet is a particular metadata field that is considered important for the data set that is being navigated. By selecting a particular facet value in the context of browsing, Seamark adds a facet restriction and removes all items that do not meet that restriction. The browser window is then updated to reflect the new subset of relevant data. Selecting facet values has the effect of zooming in on the data set, by removing links to data that are no longer of interest. With faceted navigation, it is possible to remove a restriction that was made at an earlier point, thereby zooming out to increase the field of search. Seamark has a flexible administrative environment that allows customized interfaces to be easily designed, and for different views to be published to different individuals. Users can also perform traditional keyword search for data within a whole RDF graph or a pre-selected subset.

Seamark Navigator has been integrated with the Oracle RDF Data Model to enable the effective search and navigation of RDF graphs within the Oracle Database. In the current integration, the RDF graph is extracted from the Oracle RDF Data Model into Seamark using the Oracle graph query capability. Once the data are loaded into Seamark, the data are indexed to support the generation and execution of faceted browsing over the data. A more extensive level of integration is underway that will enable data to remain within the Oracle RDF Data Model at query time, thus enabling Seamark to take full advantage of the scalability features of the Oracle Database.

3.0 Drug Discovery Use Case

The use case is provided to demonstrate the application of the Oracle RDF Data Model and Seamark Navigator to data search and browsing within drug discovery. The example aggregated several publicly available bioinformatics data sets into the Oracle RDF Data Model and utilized the Seamark Navigator interface for the exploration, organization and visualization of these data within the RDF infrastructure. A gene expression data set was selected to exercise the RDF infrastructure.

3.1 Overview

The use case is derived from research originally undertaken by Shipp et al. [19] where gene expression microarray experiments were used to study the disease characteristics of patients with diffuse large B-cell lymphoma and their response to anticancer treatment. These experiments provide a basis whereby genes can be identified that would differentiate between the different forms of disease and lead to the identification of patient subgroups. The classification of patients into subgroups can assist with the discovery and development of more targeted treatments for individuals. Once gene biomarkers have been identified, it is necessary to determine their function, biological and chemical properties, disease associations, and role in biological pathways.

The following sections describe the implementation of an RDF infrastructure for the interrogation of probesets of interest, and the results gained from that exploration.

3.2 Data Mining

Oracle Data Mining (ODM) was used to identify top biomarker genes for a subset of patients with diffuse large B-cell lymphoma that do not respond to chemotherapy. The raw gene expression measurements from an Affymetrix scanner were loaded into the Oracle Database using the SQL*Loader utility. The Minimum Description Length algorithm for determining attribute importance in ODM [8] identified 88 values with positive influence on the outcome. These 88 probesets provided information that could help distinguish those patients who do not respond to chemotherapy, from those patients that do. As little insight can be gleaned from probeset identifiers alone, the RDF infrastructure was used to identify biologically interesting information relating to the 6 probesets with the highest importance values (Table 1).

Table 1. Top Biomarker Genes for Differentiating Patients with Diffuse Large B-Cell Lymphoma.

Affymetrix Probeset	**Rank**	**Importance**
M17863_s_at	1	0.12034767
L28175_at	2	0.09065704
L40377_at	3	0.08976085
J05036_s_at	4	0.08450493
U43519_at	5	0.08120555
M18255_cds2_s_at	6	0.07303495

3.3 Data Exploration and Results

The top six probesets from the gene expression analysis were entered into Seamark Navigator in order to retrieve related gene and protein information, including Affymetrix description, Entrez GeneID, UniProt Keywords, and Gene Ontology (GO) ID.

It was discovered that one of the probesets that was ranked highly by the Minimum Description Length algorithm was for the gene Protein Kinase C Beta. This was of interest because Protein Kinase C is known to be a critical protein messenger in the transfer of growth signals for B-cells and B-cell lymphomas [19].

By selecting all six probesets of interest simultaneously for further drill down, it was determined that they corresponded to 43 GO terms. It was further revealed that there was clustering within the GO molecular function classes of receptor activity, receptor binding, hydrolase activity, and transferase activity. The BiNGO plugin [11] for Cytoscape [18] was used to undertake a Binomial test with Benjamini & Hochberg Fales Discovery Rate correction to determine that the clustering was significant to $P < 0.05$.

3.4 Infrastructure Implementation

To provide support for the RDF data exploration through Seamark Navigator, twelve publicly available bioinformatics data sets were identified. These collectively contained a wide range of biologically relevant data. Each data set was manually examined to identify a cross reference that would be needed in order to map between the different data sources. The goal was to create a concept map that linked all of the biological entities to one another, enabling users to easily jump to information of interest among the different data sets. Several additional data sets were required in order to achieve all of the desired mappings, for example, gene2go and ec2go. The interconnectivity of the chosen data sets is shown in Figure 1.

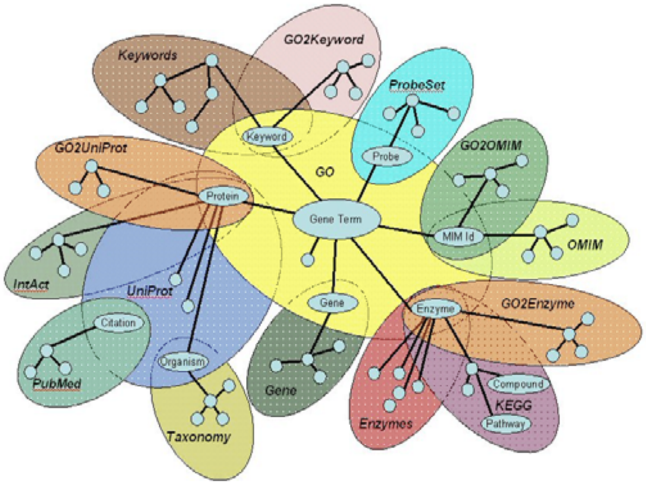

Figure 1. Linkage Points between the Selected Bioinformatics Data Sets.

URIs were assigned to the biological entities within the data sets. The Life Sciences Identifier (LSID) standard was used in data sets that support this standard [4]. In other instances, data set proprietary unique identifiers were used to generate URN identifiers.

Enzymes, GO, IntAct, NCBI Taxonomy, and UniProt were already available in RDF/XML, so these data sets were simply downloaded from the Web. For the data that was only available in a flat file format, the XML DB feature within the Oracle Database was used to convert data in a tab-delimited format into XML. The data were either loaded into the database for the transformation, or accessed externally through Java Database Connectivity (JDBC). Extensible Stylesheet Language Transformations (XSLT) were then applied to each of the data sets in XML to convert them all into RDF/XML.

The whole of GO was downloaded for the use case, but in all other instances sub-sets of data were used. In total, the data generated 316,296 triples (http://www.olsug.org/wiki/index.php?title=Image:JWS.zip). The focus of the work was to explore the interoperability among many different life sciences data sources, as scalability has been previously examined. Further details regarding the data sets used, and the structure of the URIs is provided in Table 2.

Once the RDF data were loaded into the Oracle RDF Data Model, rules were used to link the data sets together. If different data sets co-referenced the same URN, then the data regarding that particular entity was collapsed. In this use case, all data mappings were manually examined to ensure correctness. If the use case were to be deployed in a production environment, it would be possible to write a program that could automatically perform the mappings between the data sets.

Table 2. Bioinformatics Data Sources Utilized in the Case Study

Data Set	Data Set Content	URI Origin	URI Example
Affymetrix Probesets	Probesets	Created	urn:lsid:uniprot.org: probe:J05036_s_at
Entrez Gene	Genes	Created	urn:lsid:uniprot.org: gene:814642
Enzymes	Enzymes	Created	urn:lsid:uniprot.org: enzymes:1.16.1.7
Gene Ontology	Ontology	LSID	urn:lsid:uniprot.org: go:3674
IntAct	Protein Interactions	LSID	urn:lsid:uniprot.org: intact:EBI-367757
KEGG	Compounds	Created	urn:lsid:uniprot.org: pathway:map00190
KEGG	Pathways	Created	urn:lsid:uniprot.org: compound:C00003
NCBI Taxonomy	Organisms	LSID	urn:lsid:uniprot.org: taxonomy:12333
OMIM	Diseases	Created	urn:lsid:uniprot.org: omim:106195
PubMed	Literature	LSID	urn:lsid:uniprot.org: pubmed:15143089
UniProt	Proteins	LSID	urn:lsid:uniprot.org: uniprot:Q62225
UniProt	Keywords	LSID	urn:lsid:uniprot.org: keywords:2

As Seamark Navigator is a faceted browser, it was necessary to select the specific facets for each data set. In order to ensure an intuitive user experience, attention was paid to which facets were retrieved at each stage during the browsing process. The initial Web page interface was designed to assist the user in identifying the data that they would be interested in interrogating. Subsequent Web pages were designed to help guide the user to relevant information of interest, while using filters to minimizing the size of the RDF search space. The interface was designed to enable users to either retrieve data about a single biological entity, or to retrieve data that applied to a group of entities. Once the facets had been determined, Seamark was used to generate a faceted browsing interface.

4.0 Discussion

Semantic Web standards have matured to a point where commercial software solutions are available to address real-world problems. This paper provides insight into the power of RDF, and the use of the Oracle RDF Data Model and the Seamark Navigator from Siderean to the infrastructure for a biological use case.

The Oracle RDF Data Model has generated much attention. This stems from the desire to be able to manage RDF data in a secure, scalable, and highly available environment. Users have the flexibility of being able to incorporate multiple media data, such as images and text, into the RDF graphs. It also provides the ability to perform queries that span the three common data formats: relational, XML, and RDF.

The ability to effectively manage data in RDF within the Oracle Database simplified the process of making data available to Seamark Navigator. This is because it is now no longer necessary for Seamark to perform transformations upon the data to make it available in RDF. Additional benefits include simplified integration of RDF data with other enterprise data, re-use of RDF objects, eliminating modeling impedance mismatch between client RDF objects and relational storage, and easier maintenance of RDF applications.

With the current level of integration, the Oracle graph query is used to pass requested RDF data to Seamark. However, work is underway to move the data query into the Oracle Database in order to provide a more scalable solution for large data sets. Currently, a single instance of Seamark Navigator operating on a 32-bit processor provides support for up to approximately 20 million triples, and a clustered implementation can scale up to hundreds of millions of triples. The integration of Seamark Navigator with the Oracle RDF Data Model is sought to maintain low latency while achieving an order of magnitude higher scalability.

As data volumes continue to grow in the life sciences, it becomes increasingly important to have effective mechanisms for browsing data and to be able to query subsets of data of interest. Faceted navigation helps overcome some of the limits of search by revealing the scope of information available. This provides context for exploration, discovery, and selection of the content that is most valuable. Seamark eliminates the need to know in advance what is stored in the data repository and how it is classified.

The RDF infrastructure deployed was able to easily and quickly retrieve biological data that related to the probeset biomarkers. It also became possible for users to select or search for the bioinformatics information of interest, rather than manually collecting identifiers to enable jumping between data sources. In addition, the ability of faceted browsing to remove all data that does not meet filter conditions, and being able to select multiple entities for simultaneous search, it was possible to identify clustering of genes within GO. In this use case, it appears that the results may be of biological interest, as the clustering was found to be of statistical significance by the BiNGO plugin for Cytoscape.

However, the identification of optimal facets for browsing can be difficult when there is a large number of instances assigned to a particular facet, as frequently occurs in the life sciences. For example, thousands of entities would be under the facet heading of genes. In addition, faceted browsing does not provide all necessary analytical tools and visualization capabilities required by bioinformatics.

The infrastructure highlighted in the paper takes advantage of RDF, with the goal of showing how this layer alone can be used to successfully aggregate data. This approach was decided upon as many life sciences organizations wish to achieve data integration with ultimate flexibility. However, RDFS and OWL can be used to provide a common vocabulary and a more formal framework for data integration.

The availability of the Oracle RDF Data Model and Seamark Navigator aided the building of the RDF infrastructure. However, these products did not obviate the need to identify data sets that were cross-referenced, the transformation of data sets into RDF, and the assignment and mapping of unique identifiers. Data sets were selected such that they covered many biological data types of interest (e.g. genes, proteins, disease), and had sufficient cross-references to enable all of the data to be connected. It was therefore necessary to have a good understanding of the data sets, and how they interrelated. The transformation of data into RDF was relatively straightforward.

LSIDs were used to generate the URIs in data sets that supported this convention. However, in some cases it was necessary to use different identification schemes. In order to link LSIDs with data set proprietary identifiers, it was necessary to undertake manual linking of data sets. Going forward, it is hoped that the life sciences community agrees upon a convention for the identification of entities, as this will enable biological data sets to be cross-referenced far more easily.

However, assigning unique identifiers is not a simple task within the life sciences, as data are frequently described at different semantic levels. For example, in articles about gene expression analysis it is common for a gene name such as PRKCB1 to be referenced, rather than the probeset M18255_cds2_s_at. In articles about protein interaction it is common to state that two genes interact, whereas it is really meant that the proteins produced by the genes interact. This can lead to incorrect biological assumptions, as genes typically produce a variety of splice variants, and some of these variants will almost certainly not interact. There is also ambiguity as to how proteins complexes are referenced in literature. For example, it may be stated that a protein interacts with a second protein, when, in reality, both protein names consist of a complex of proteins. Consideration, therefore, needs to be taken when identifiers are assigned to ensure that they are at a level of granularity that will help biological understanding to be furthered. In the case of protein identifiers, it would be valuable if references pointed to specific proteins, their molecular state, and other information that would influence their behavior.

Much progress has been made in the adoption of LSIDs for proteins and genes, and INChI [5] for chemical compounds. However, over time, methods need to be developed for assigning identifiers to protein complexes, protein interactions, and pathways. Currently, as a consequence of the ambiguity in naming of entities, care must be taken to ensure equivalence when data sets are merged. Further, much of what is considered equivalence can depend on the semantic context of the operation.

Many technology approaches have been used for data integration projects. Data warehouse approaches have been built to enable users to consolidate data into a relational database environment. Challenges with this approach include the substantial planning and upfront effort required to unify the many data sources, ongoing efforts are required to maintain the data model and loaders, and a limited ability to share relational data with other organizations. However, relational data stores do provide a highly available, scalable, and secure environment. An alternative approach is federated data access, which allows users to read or update data in multiple distributed data stores as if the data were loaded into a single database. All vendors of enterprise relational databases offer such solutions. A challenge with a federated database approach is that it does not require consistency between the data models that are being queried. As any incompatibilities between the systems are discovered, for example, naming conventions or units of measure, these problems will need to be rectified and changes made to the applications that access the data.

Many organizations are exploring the use of Web services for defining domain-specific ways of describing information and exchanging those definitions between applications. This enables information to be viewed in context, although the application environment is distributed. However, there are some challenges to the use of Web services in drug discovery. For example, the life sciences community has not agreed upon a single set of XML schemata, which limits the flexibility of the approach, and the dynamic nature of scientific data requires XML schemata to be updated on a frequent basis [22]. In addition, while Web services allow the correct syntactic integration by passing valid instance data between scientific application systems, the semantics are not defined, thus relying on the knowledge of the person integrating the applications to know the meaning of the low-level data structures. However, a number of research projects are focused on applying semantics to Web services [12, 13, 21].

The Semantic Web has the benefit of helping to ensure that two concepts that are found in different forms in different data sources actually describe the same object. Being able to recognize homonyms, synonyms, and related terms is critical in data integration. RDF provides a very flexible data model for adding new data to both individual data sets and knowledge banks. The Semantic Web provides standard specifications, such as OWL, that can assist with uniting data if differing identifiers were initially selected. This functionality is especially valuable when different departments choose to aggregate their data, for example bioinformatics with cheminformatics.

5.0 Summary

The integration of biomedical data within drug discovery has proven to be a long-standing challenge. Semantic Web technology promises the ability to more easily aggregate such data, thereby improving the efficiency of drug discovery. The Semantic Web infrastructure deployed enabled many disparate life sciences data sets to be integrated. The Oracle RDF Data Model provided a secure, scalable, and highly available environment for managing the data. Seamark Navigator provided an effective environment in which to explore the relationships within the data, and drilldown in areas of interest.

Acknowledgements

We are grateful to the Oracle Spatial group for implementing the RDF Data Model, Bob Haberstroh for guidance with the data mining, and Melissa Cline and Ora Lassila for feedback on the manuscript. JL wishes to thank the National Science Foundation for support (IIS-0542041).

References

[1] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American, 284 (2001) 34-43.

[2] K.-H. Cheung, K.Y. Yip, A. Smith, R. deKnikker, A. Masiar, M. Gernstein, YeastHub: a semantic web use case for integrating data in the life sciences domain, Bioinformatics 21 (2005) i85-i96.

[3] E. Chong, S. Das, G. Eadon, J. Srinivasan, An efficient SQL-based RDF querying scheme, Proceedings of the 31st international conference on very large databases, Trondheim, Norway. (2005) 1216-1227.

[4] T. Clark, S. Martin, T. Liefeld, Globally distributed object identification for biological knowledgebases, Briefings in Bioinformatics 5 (2004) 59-70.

[5] S.J. Coles, N.E. Day, P. Murray-Rust, H.S. Rzepa, Y. Zhang, Enhancement of the chemical semantic web through the use of InChI identifiers, Organic and Biomolecular Chemistry 3 (2005) 1832-1834.

[6] S. de Coronado, M.W. Haber, N. Sioutos, M. S. Tuttle, L. W. Wright, NCI thesaurus: using science-based terminology to integrate cancer research results, Medinfo 2004 (2004) 33–37.

[7] J. Goldbeck, G. Fragoso, F. Hartel, J. Hendler, B. Parsia, J. Oberthaler, The national cancer institute’s thesaurus and ontology, Journal Web Semantics 1 (2003) 1-5.

[8] I. Kononenko, On Biases in estimating multi-valued attributes, in: C. Mellish, ed., Proceedings International Joint Conference on Artificial Intelligence, Montreal, Canada, 223 (1995) 1034-1040.

[9] A. Kozlenkov, M. Schroeder, PROVA: rule-based java-scripting for a bioinformatics semantic web, in: E. Rahm, ed., Proceedings of Data Integration in the Life Sciences, 1st International Workshop, DILS 2004, Leipzig, Germany, March 2004, Springer-Verlag, Lecture Notes in Computer Science 2994 (2004) 17-30.

[10] J. Luciano, PAX of mind for pathway researchers, Drug Discovery Today 13 (2005) 938-942.

[11] S. Maere, K. Heymans, M. Kniper, BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks, Bioinformatics 21 (2005) 3448-3449.

[12] S. Majithia, D.W. Walker, W.A. Gray, Automating scientific experiments on the semantic grid, In: S.A. McIlraith, D. Plexousakis, F. van Harmelen eds., 3rd International Conference, The Semantic Web – ISWC 2004, Hiroshima, Japan, November 2004, Springer-Verlag, Lecture Notes in Computer Science 3298 (2004) 365-379.

[13] S.A. McIlraith, T.C. Son, H. Zeng. Semantic Web Services, IEEE Intelligent Systems 16 (2001) 46–53.

[14] F. Manola, E. Miller, RDF Primer. W3C Recommendation, 10 February 2004. http://www.w3.org/TR/rdf-primer/.

[15] D.L. McGuinness, F. van Harmelen, OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/.

[16] E. Neumann, E. Miller, J. Wilbanks, What the semantic web could do for the life sciences, Drug Discovery Today: Biosilico 2 (2004) 228-236.

[17] E. Prud’hommeaux, A. Seaborne, SPARQL Query Language for RDF. W3C Candidate Recommendation 6 April 2006. http://www.w3.org/TR/rdf-sparql-query/.

[18] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, A. Amim, B. Schwikoski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Research 13 (2003) 2498-2504.

[19] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nature Medicine 8 (2002) 68-74.

[20] S.M. Stephens, Enabling semantic web inferencing with Oracle technology: applications in life sciences, in: A. Adi, S. Stoutenburg, S. Tabet, eds., Proceedings of 4th International Workshop, RuleML 2005, Galway, Ireland, November 2005, Springer-Verlag, Lecture Notes in Computer Science 3791 (2005) 8-16.

[21] R.D. Stevens, H.J. Tipney, C.J. Wroe, T.M. Oinn, M.Senger, P.W. Lord, C.A. Goble, A. Brass, M. Tassabehji, Exploring Williams-Beuren syndrome using myGrid, Bioinformatics 4 (2004) I303–I310.

[22] X. Wang, R. Gorlitsky, J.S. Almeida, From XML to RDF: how semantic web technologies will change the design of ‘omic’ standards, Nature Biotechnology 23 (2005) 1099–1103.

Susie Stephens is a Principal Product Manager at Oracle, where she is responsible for enhancing core technology products to make them more suited to the needs of the life sciences industry. She was heavily involved in the implementation of the Oracle RDF Data Model, guides Oracle in its adoption of the Semantic Web, and represents Oracle in discussions with W3C in this area.

David LaVigna heads Siderean Software’s Customer Services organization where he is responsible for implementation and customer satisfaction. He brings to Siderean more than fifteen years experience working in the IT and software development spheres.

Mike DiLascio heads business development for Siderean Software. Over a twenty-two year career in the computer software industry, he has specialized in commercializing emerging technology while delivering measurable value and operational improvements to Global 1000 and government customers.

Joanne Luciano is a Lecturer of Genetics at Harvard Medical School, an Honorary Visiting Research Fellow in the Department of Computer Science at the University of Manchester, England, and the President and Founder of Predictive Medicine, Inc. She is an active leader in BioPAX, the BioPathways Consortium, and W3C’s Healthcare and Life Sciences interest group.