Provenance of Microarray Experiments for a Better Understanding of Experiment Results

Helena F. Deus
Department of Bioinformatics and Computational Biology
The University of Texas M. D. Anderson Cancer Center
Houston, USA
Instituto de Tecnologia Química e Biológica, UNL
Lisboa, Portugal

Jun Zhao
Deparment of Zoology
University of Oxford
Oxford, UK

Satya Sahoo
Kno.e.sis Center
Department of Computer Science and Engineering
Wright State University
Dayton, USA

Mathias Samwald
Digital Enterprise Research Institute
National University of Ireland Galway
Galway, Ireland

Eric Prud'hommeaux
World Wide Web Consortium
MIT
Cambridge, USA

Michael Miller
Tantric Designs
Seattle, USA

* M.Scott Marshall
Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
Leiden, The Netherlands
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands

* Kei-Hoi Cheung
Center for Medical Informatics
Yale University School of Medicine
New Haven, USA

***Abstract*—This paper describes a Semantic Web (SW) model for gene lists and the metadata required for their practical interpretation. Our provenance information captures the context of experiments as well as the processing and analysis parameters involved in deriving the gene lists from DNA microarray experiments. We demonstrate a range of practical neuroscience queries which draw on the proposed model. Our provenance representation includes the origins of the gene list and basic information about the data set itself (e.g. last modification date and original data source), in order to facilitate the federation of gene lists with other types of Semantic Web-formatted data and include the integration of a broader molecular context through additional omics data.**

Keywords-data integration, query federation, semantic web

I. INTRODUCTION

In the genomics/post-genomics era, massive amounts of data generated by high throughput experiments, including those using microarray technologies, have presented both promises and challenges to clinical, and translational research. One goal of microarray experiments is to discover, out of tens of thousands of genes, a small subset of genes (usually on the order of hundreds) whose expression pattern is indicative of some biological response to a given experimental condition.

Many computational/statistical approaches have been developed to detect such biologically significant gene lists. According to [1], the workflow of a microarray experiment is divided into the following steps: i) **experimental design** that includes the type of biological questions the experiment is designed to address, how the experiment is implemented (e.g., experiment and control), sample preparation, microarray platform selection, hybridization process, and scanning; ii) **data extraction**, which includes image quantification, filtering, and normalization; and iii) **data analysis and modeling**, which include approaches such as clustering, t-tests, enrichment analysis and so on.

The gene lists produced in step iii are usually reported as part of the experimental results published in scientific papers, and the steps involved in obtaining the gene lists are described in the methods section. Sometimes, gene lists are made electronically available (e.g., spreadsheets) through journal web sites. However, to the best of our knowledge, there is no standard format for uniformly representing and broadly sharing such gene lists in a focused scientific context.

We believe it would be useful to the community if such gene lists were commonly represented in a standard SW vocabulary and accessible to SW applications. This approach makes it possible for researchers to work with the gene list without requiring a post hoc significance analysis to re-derive the list. If experimental factors are included with gene lists, researchers can account for context without requiring labor-intensive manual research into the experimental factors for

*These authors contributed equally to this work. KC is supported in part by NIH grant U24 NS051869. JZ is supported by EPSRC grant EP/G049327/1. HFD is supported by the portuguese FCT (Fundação para a Ciência e Tecnologia) scholarship SFRH/BD/45963/2008. The work of MS was funded by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2) and by a postdoctoral fellowship from the Konrad Lorenz Institute for Evolution and Cognition Research, Austria.

each microarray study. A standard representation can be used both for gene lists reported in individual papers (note that these published gene lists are not yet stored in most microarray databases) and those computed from datasets collected from multiple microarray experiments across different microarray databases (e.g., GEO profiles [2] and Gene Expression Atlas [3]).

Integrated analysis (meta-analysis) requires raw and processed datasets from independent microarray experiments to be selected, compared, combined, and correlated using a variety of computational/statistical methods. This is, of course, much easier with machine-readable provenance and experimental context. To this end, MIAME [4] was proposed by the Microarray Gene Expression Data (MGED (http://www.mged.org)) community (now called "Functional Genomics Data Society" or FGED) to describe the *Minimum Information About a Microarray Experiment* (MIAME) that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. MIAME represents a set of guidelines for microarray databases and data management software. The MAGE data model and MAGE-ML (a standard XML format for serializing the MAGE model) [5] have been developed based on the MIAME data content specifications. In addition, MAGE-TAB [6] was proposed as a (more user-friendly) alternative to MAGE-ML.

Along with the development of these standards, a significant number of microarray databases ranging from individual labs (e.g., Nomad at deRisi lab (http://ucsf-nomad.sourceforge.net/)), institutions (e.g., SMD 7], YMD [8], and RAD [9]) to the scientific community (e.g., GEO [2] and ArrayExpress [10]) have been created, making large collections of microarray datasets accessible to the public. There are also microarray databases that serve the needs of specific biomedical domains (e.g., the NIH Neuroscience Microarray Consortium (http://np2.ctrl.ucla.edu/np2/home.do)). Major journal publishers have promoted sharing of microarray data by requiring authors to submit their data to public microarray repositories. Some journal publishers make supplemental data available on their web sites.

While many microarray databases are MIAME-compliant, several challenges still remain for researchers wishing to locate datasets relevant to their interest:

- There is no central repository for all microarray datasets, and experiment/dataset are stored on multiple databases.
- Users must learn to use different search interfaces and analytic facilities at each database.
- Many databases lack experimental context, annotation, and provenance.
- There is a lack of use of standard vocabularies in many microarray databases.
- The lists of differentially expressed genes discussed by most articles associated with a microarray study are not disclosed in any standard format, nor are they programmatically accessible.

The Semantic Web [11] has been actively explored in the context of biomedicine. For example, the W3C Semantic Web Health Care and Life Sciences Interest Group (HCLS IG) (http://www.w3.org/2001/sw/hcls/) represents a major community effort involving both academia and industry. The HCLS IG and allied efforts provide a growing corpus of biomedical datasets expressed in the Resource Description Framework (RDF) and web ontology language (OWL). Wang et al [12] has described how the transition from the eXtended Markup Language (XML) to RDF could potentially enhance semantic representation and integration of omic data. In addition to data, biomedical ontologies are made available to the community through organizations such as NCBO (http://www.bioontology.org/) and OBO Foundry (http://www.obofoundry.org/).

In this paper we explore using SW to represent microarray experimental data and provenance information about the context under which the data were generated, including the goal of the experiment, experimental factors (such as the disease or the cell region), and the statistical analysis process which leads to the experiment results. We explore the role of provenance information in helping biologists understand microarray experiments in the context of other experiments as well as other existing biomedical knowledge. To facilitate a quality-aware federation of microarray experiment results, we also provide provenance information about the gene lists data published using SW standards. As a pilot study, we take a bottom-up approach focusing on the type of provenance information required to meet our motivation use cases and creating a representation model with the minimum set of terms to meet these use cases. Although these terms are currently defined in our own namespaces, they can largely be mapped to existing provenance vocabularies, which are generically defined and evolving, to achieve maximum interoperability, in the next stage of our pilot study.

II. MOTIVATION

One motivation of microarray experiments is to identify genes that are differentially expressed in biological samples under different conditions (e.g., disease vs. control). The samples may come from tissues extracted from different organs or parts of the same organ (e.g., different brain regions). In this case, we may be able to discover differentially expressed genes in each organ/organ part and how disease may affect each organ/organ part at the gene expression level. A common outcome of experiments is a list of candidate genes which may serve as diagnostic or therapeutic markers. These gene lists, abundant in biomedical literature, are provided in heterogeneous formats (e.g., Excel spreadsheets and printed tables embedded in papers) that hinder the reuse of the results. In order to reuse such gene lists in additional pathway or molecular analysis, it is important that they are represented in a standardized, distributable, and machine-readable format that is amenable to semantic queries.

After obtaining a representative list of differentially expressed genes, scientists may need to study these experiment results in a broader molecular context with

additional data. In the case of neurological disease studies such as Alzheimer's Disease (AD), researchers may want to combine gene expression data from multiple AD microarray studies. For example, one characterization of AD is the formation of intracellular neurofibrillary tangles that affect neurons in brain regions involved in the memory function. It is important to have meta-data such as the cell type(s), cell histopathology, and brain region(s) for comparing/integrating the results across different AD microarray experiments. It is important also to consider the (raw) data source and the types of analysis performed on the data to arrive at meaningful interpretations. Finally, gene expression data may be combined with other types of data including genomic functions, pathways, and associated diseases to broaden the spectrum of integrative data analysis.

In our pilot study, we selected three microarray experiments from different journals ([13-15]) to explore how to represent gene list experiment results in a structured format and what types of metadata can better enable the computer to search for genes that may play a molecular role in the pathogenesis of AD. All the gene lists from the selected publications were derived from human brain samples that were prepared for AD studies. We wanted to be able to answer a variety of user questions regarding semantically related experiments and their experimental results. For example:

- Q0: What microarray experiments analyze samples taken from the Entorhinal cortex region of Alzheimer's patients?
- Q1: Was the same data normalization algorithm or statistical software package used in both studies that analyze gene expression in the entorhinal cortex region of AD patients?
- Q2: What genes are overexpressed in the Entorhinal cortex region in the context of Alzheimer's and what is their expression fold change and associated p-value?
- Q3: Are there any genes that are expressed differently in two different brain regions (such as in Hippocampus and Entorhinal cortex)?

The MIAME standard outlines the minimum set of information that is needed for describing microarray experiments in order to facilitate the reproduction of these experiments and a uniform interpretation of experiment results. Experiments recording and publishing MIAME-compliant experimental protocol should contain sufficient information to answer questions like Q0 and Q1. However, because MIAME does not specify a format, and MAGE-ML and MAGE-TAB do not specify a standard representation for experiment results (such as the set of genes showing particular expression patterns), there is no simple mechanism to find semantically related experimental results based on the patterns of differentially expressed genes.

In order to answer questions Q2 and Q3, it is necessary to model both experimental information (ex: Entorhinal cortex) and statistical data (e.g. the p-values associated with gene expression values).

Additionally, we want to be able to extend the knowledge about genes linked to AD such that scientists can access and extend their understandings about their gene expression data analysis results to answer questions like the following:

- Q4: What other diseases may be associated with the same genes found to be linked to AD?
- Q5: What drugs are known that affect the same overexpressed gene products and what are their target diseases?
- Q6: Select all the genes determined to be differentially expressed in the Entorhinal cortex in experiments performed by AD investigators at the Translational Genomics Research Institute

For these types of questions, the microarray experiment results need to be federated (Q4, Q5) or combined (Q6) with other datasets describing the data itself. We show how the structured representation of microarray experiment data and associated provenance metadata will enable us to query across different aspects of domain knowledge about these experiment results using several other datasets in the HCLS KB. We also show how we can provide additional provenance information about different datasets to support some quality-aware federation queries over distributed data sources.

III. METHODS

To address questions Q0-Q3 we need both a precise representation of the gene lists reported in the three selected publications and a representation of the provenance of these gene lists, such as the methods and procedures involved in their generation. As mentioned in Section I, several standards exist for describing microarray experiment protocols, however, none is comprehensive enough to fully capture the complex process of reporting the results of a microarray experiment. To answer questions Q4-Q5 we need to query across the exemplar datasets, using provenance information of different levels of granularity, from the basic information about the context of each experiment to details about the analysis processes generating the gene expression results. Although a number of provenance vocabularies, such as the open provenance model (OPM, http://openprovenance.org/) and Provenir (http://wiki.knoesis.org/index.php/Provenir_Ontology) are available, we choose a bottom-up approach in this pilot study. On the one hand, at the time of the writing, little was known about how to choose between these existing vocabularies to best suit our purpose; on the other hand, our pilot study aims to focus on capturing the minimum information to answer our case study questions. This approach has the added advantage of shielding our model from having to keep pace with rapidly evolving ontologies while still enabling mapping to upper level ontologies in the future. For these reasons, our data model includes the minimum set of terms necessary to describe the three examples selected, and is made available under our own local namespace:

@prefix biordf:http://purl.org/net/biordfmicroarray/ns#

Compared with provenance vocabularies, many domain specific ontologies are much more established and stable, such as NIF (http://www.neuinfo.org/), disease ontology (DO, http://do-wiki.nubic.northwestern.edu/index.php/Main_Page), or the voiD vocabulary [16]. Therefore, we reuse terms from

these ontologies that are already widely used to annotate (biological) datasets in our data model in order to enable maximum interoperability with other approaches.

A. *The Data Model*

Our data model captures the minimum information necessary to describe the gene lists and the microarray experiment context in which they were generated. To answer each of the individual case study questions, different aspects of each dataset had to be considered. For example, to answer questions like **Q0** and **Q3** a good overview of each microarray experiment is necessary, including the samples used, the disease of interest, microarray platform, etc. For questions like **Q1** and **Q2**, however, a different set of assertions concerned specifically with comparing gene expression quantification methods in different settings is required. Finally, the ability to answer questions like **Q4** and **Q5** involve the more complex component of performing simultaneous queries on more than one data source. As such, information describing the metadata associated with each data source is also necessary. To accommodate these different data types in our model, we have defined four provenance levels, with each level entailing different subsets of information:

Institutional level: Includes assertions about the laboratory where the experiments were performed and the reference where the results were published to help determine the trustworthiness of the data. This information is useful to constrain the list of significant genes to only those that are published in peer-reviewed articles and/or were performed at certain institutions that have the track record of generating high quality microarray data published in respected journals.

Experiment protocol level: Includes assertions about the brain regions from which the samples were gathered and the histology of the cells. Such information has been partially mapped to MGED, DO and NIF terms.

Data analysis and significance level: Includes assertions about the statistical analysis methodology for selecting the relevant genes. Terms defined for this level are also provided as a separate statistic module (http://purl.org/net/biordfmicroarray/stat#) to describe software tools and statistical terms.

Dataset description level: Includes assertions about when the dataset is published, based on which version of a source dataset, and who published the dataset. Some existing vocabularies for describing RDF datasets on the Web were reused to enhance their trustworthiness such as the Vocabulary of Interlinked Dataset (voiD) [16] that provide basic information about who published the data as well as a summary of the content of the dataset, such as the number of genes described by the dataset or the SPARQL endpoint through which the dataset can be accessed. The Provenance Vocabulary [17] was also used to provide a richer set of provenance information, such as when the dataset is published, using which tool, or by accessing which data server.

B. *Formulation of SPARQL queries*

The queries described here are formulated at our demo site (http://purl.org/net/biordfmicroarray/demo), where they can be directly executed or copied and performed locally using software such as SWObjects (https://sourceforge.net/projects/swobjects/files/). The demo site also includes a diagram explaining the four provenance levels and the types of data entailed in each level.

To answer **Q0**, experiments performed in samples collected from patients with Alzheimer's disease in a specific area of the brain, the Entorhinal cortex, must be selected from the RDF representation. The data necessary to answer to this question is completely entailed in the experimental provenance level and can be formulated in terms of the entities used to represent each step of the workflow involved in collecting a Sample. Making use of data from the statistical analysis provenance level, the same query **Q0** can be amended to filter the list of experiments retrieved based on the statistical normalization software thus enabling an answer to **Q1**. To answer questions **Q2** and **Q3** data pertaining to the experiment provenance level must also be combined with information about the gene lists, such as the expression level for each gene. A common requirement to measure statistical significance of differentially expressed genes is the p-value that is associated with gene expression fold change. In **Q2**, this information is used to trim the list of over-expressed genes by indicating that fold change > 0 but only in cases where the p-value is < 0.001.

One of the most significant advantages of representing gene lists in RDF is helping scientists enrich it with data from linked datasets such that questions like **Q4** and **Q5** may be answered. The dataset description provenance level enables the discovery of useful datasets for specific purposes, such as, e.g. using the HCLS Kb to discover diseases that may be associated with specific genes. **Q4**, detailed below, achieves that goal by first retrieving the same list of genes as in **Q2** and, secondly, by selecting the most recently updated SPARQL service which includes assertions about both genes and diseases. The final section queries this service to retrieve the correlated diseases.

```
SELECT DISTINCT  ?diseaseName ?geneLabel ?geneName WHERE {
#Retrieve a list of overexpressed genes in the entorhinal cortex of AD patients
{
  ?experimentSet dct:isPartOf ?microarray_experiment ;
                 biordf:has_input_value ?sampleList ;
                 biordf:differentially_expressed_gene ?gene ;
                 biordf:has_ouput_value ?foldChange .
  ?sampleList  biordf:derives_from_region ?brainRegion ;
               biordf:patients_have_disease ?alzheimers .
  ?gene  rdfs:label ?geneLabel ;
         biordf:name  ?geneName .
  ?foldChange rdf:value ?foldChangeValue ;
              stat:p_value ?pval .
  #Apply filters to constrain the amount of results
    FILTER (xsd:float(?foldChangeValue) > 0)
    FILTER (xsd:float(?pval) < 0.001 )
    FILTER (?brainRegion = neurolex:Entorhinal_cortex )
    FILTER (?alzheimers = doid:DOID_10652 )
}
#Find most recently updated SPARQL endpoint that contains information about genes and diseases.
{
```

```
        ?source rdf:type void:Dataset ;
        void:sparqlEndpoint ?srvc ;
        dct:issued ?issued  ;
        dct:subject diseasome:diseases ;
        dct:subject diseasome:genes .
    OPTIONAL {
        ?source1 rdf:type void:Dataset ;
        void:sparqlEndpoint ?srvc2 ;
        dct:issued ?issued2 ;
        dct:subject diseasome:diseases ;
        dct:subject diseasome:genes .
        FILTER (?issued2 > ?issued)
    }
    FILTER (!BOUND(?srvc2))
    }
    #Get associated diseases from most recently updated Diseasome server.
      SERVICE ?srvc2 {
        ?diseasomeGene rdfs:label ?geneLabel .
        ?disease diseasome:associatedGene ?diseasomeGene.
        ?disease rdfs:label ?diseaseName .
    }
    }
```

Finally, to answer **Q6** data from the institutional provenance level we must limit the list of retrieved experiments to those that were performed at a specific institution. The queries presented here are executable through our demo at http://purl.org/net/biordfmicroarray/demo. Their time to execution ranges between 100 and 200 ms for local queries (Q1-Q3, Q6) and a few seconds (2-5s) for federated queries (Q4-Q5) executed using SWObjects.

C. *Availability*

The RDF representation was generated using JavaScript and the data was loaded into a public SPARQL endpoint (http://purl.org/net/biordfmicroarray/sparql). We elaborate and further expand the provenance queries in this paper at our demo site http://purl.org/net/biordfmicroarray/demo. A figure associating each of the four provenance levels with the data that they are concerned with is also made available at the demo site. The complete RDF/turtle representation can be downloaded from http://biordfmicroarray.googlecode.com/files/all3_genelists_provenance.ttl. The JavaScript code to convert Excel spreadsheets into RDF is available at http://code.google.com/p/biordfmicroarray/ .

IV. DISCUSSION

A data model to explicitly make the content and context of gene lists (e.g., differentially expressed genes) available in RDF format was developed. In the process, four types of provenance were identified that were found necessary to characterize, discover, reproduce, compare and integrate gene lists with other data. Expressing provenance in RDF enables describing the data itself (i.e. its origin, version and URL location) in the same language as the elements represented therein. The power of this uniform access to data and metadata should not be underestimated. In practice, this means that SPARQL queries can express constraints both about the origins of the data and contents (or attributes) of the data as demonstrated by query Q4. In the case of Linked Open Data, the set of best practices for exposing data as RDF through a SPARQL endpoint, researchers often need to distinguish between multiple RDF renderings (i.e. representations) of the same data set or different versions of it. Different endpoints can be discovered by issuing queries that target the data sources themselves: When was the last RDF rendering created and by whom (or which project)? Which ontologies/vocabularies were used? The same standardized SW mechanisms of reasoning and pattern matching can be applied to select a specific data source as the ones used to discover related facts across the data sources.

The provenance data model developed for reporting microarray experiment results while capturing different types of provenance information was motivated by our user-defined queries. We have therefore applied a bottom-up approach that focused on describing the data first before mapping it to widely used ontologies. Although several provenance ontologies are available, some of them are upper level ontologies, such as Provenir, therefore lacking the specific terms required for describing how gene lists were derived. Other ontologies, such as the Provenance Vocabulary for Linked Data and proof markup language, were created for specific application domains, such as explaining reasoning results. Our bottom-up approach enabled us to identify and define the minimum set of provenance terms to answer a set of queries from different perspectives and shield the data model from depending on external vocabularies which are often subject to changes. For increased interoperability, mapping terms from our model to terms from a community provenance model, such as the OPM or others is straightforward. For example, our property *biordf:has_input_value* can be made a sub-property of the inverse of OPM property *used*, and *biordf:derives_from_region* can become a sub-property of OPM property *wasDerivedFrom*.

Further down the pipeline of microarray studies, bioinformaticians will often need to combine knowledge about the genes derived from their microarray experiments in order to achieve a deeper understanding at a systems biology level. Although the number of genes that has to be taken into consideration while studying Alzheimer's has been significantly reduced by many gene expression studies, a good number of genes (ranging from tens to hundreds) are yet to be processed. One approach becoming increasingly popular is the use of scientific workflow workbenches (such as Taverna and Kepler) to perform large scale data analysis. Many such workbenches [19-20] also record the workflow provenance information about, for example, what genes from which organism were processed and how the proteins encoded by the genes were discovered by querying various genomic databases. Combining this workflow provenance information and the set of microarray experiment-related provenance information by mapping both to a common community provenance model, such as OPM, the trustworthiness and reproducibility of experiment results would be increased throughout the whole experiment life cycle. McCusker et al. [21] has taken a first step towards by providing a tentative translation from MGED-TAB to the OPM.

While we endorse the use of SW technologies as the standard machine-readable format, we acknowledge that most biologists are not familiar with SW and prefer to use formats such as Excel spreadsheets to work with gene list results. To this end, it would be useful to use a standardized user-friendly format (e.g., MAGE-TAB) for encoding gene lists and their context that could be easily converted into the SW format.

V. CONCLUSION

We describe and illustrate with a case study the beneficial role of Semantic Web technologies in 'omic' data representation by providing and querying a data model to capture provenance information related to reporting microarray experiment results. We have tackled not only the engineering aspect of the data integration problem, but also the more fundamental issues of federating data that begin with seemingly homogeneous data sources (microarray databases) and extends to heterogeneous data domains at multiple levels. This is also driven by the growing collaboration between a wide spectrum of scientific disciplines and communities such as is required for translational research. We have used a bottom-up approach that facilitated the identification of four provenance levels necessary to report microarray experiment results and shielded our data model from becoming dependent on constantly evolving ontologies. We have, however, discussed how some of the terms and relationships from existing provenance ontologies can be mapped to our model. Some issues found to be necessary in the integration of microarray data sources could also be considered relevant for the federation of data sources in general. As more 'omics' data are generated, the complexity and requirements for discovery-based research increases. As a result, there is a growing demand for effective data provenance and integration at many levels that counts on the active involvement of scientists and informaticians. Our work represents a step in this direction.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the HCLS IG for helping to organize and coordinate with different task forces including the BioRDF task force within which the neuroscience microarray use case was explored. Also thanks to Helen Parkinson, James Malone, Misha Kapushesky, Jonas Almeida and three anonymous reviewers. MSM appreciated the support of Jelle Goeman (LUMC) during this work.

REFERENCES

[1] Stears RL, Martinsky T, Schena M. (2003). Trends in microarray analysis. Nature Medicine. (9): 140 – 145.

[2] Barrett T, Troup DB, et al.. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009 Jan;37(Database issue):D885-90.

[3] Lukk M, Kapushesky M, et al.. A global map of human gene expression. *Nat Biotechnology* **28**, 322-324 (2010)

[4] Brazma A, Hingamp P, et al.. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 29(4):365-71.

[5] Spellman PT, Miller M, et al.. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. 3(9):RESEARCH0046.

[6] Rayner TF, Rocca-Serra P, et al.. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics. 7:489.

[7] Gollub J, Ball CA, et al.. (2003). The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res. 31(1):94-6.

[8] Cheung KH, White K, et al.. (2002). YMD: a microarray database for large-scale gene expression analysis. Proc AMIA Symp. 2002:140-4.

[9] Manduchi E, Grant GR, et al.. (2004). RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. Bioinformatics. 20(4):452-9.

[10] Parkinson H, Sarkans U, et al.. (2005). ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 33(Database issue):D553-5.

[11] Berners-Lee T, Hendler J, Lassila O. (2001). The Semantic Web. Scientific American. 284(5):34-43

[12] Wang X, Gorlitsky R, Almeida JS. (2005) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol. 23(9):1099-103.

[13] Dunckley T, Beach TG, et al.. (2006). Gene expression correlates of neurofibrillary tangles in Alzheimer's disease. Neurobiol Aging;27: 1359-71.

[14] Liang WS, Dunckley T, et al.. (2007). Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiol Genomics 28: 311-22.

[15] Liang WS, Reiman EM, et al.. (2008). Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. Proc Natl Acad Sci U S A l2008;105: 4441-6.

[16] Alexander K, Cyganiak R, Hausenblas M, and Zhao J. Describing linked datasets. In *Linked Data on the Web Workshop in the International World Wide Web Conference*, Madrid, Spain, 2009 .

[17] Hartig O, Zhao J. Publishing and consuming provenance metadata on the web of linked data. In *Proceedings of The third International Provenance and Annotation Workshop*, Troy, NY, U.S.A, 2010. In press

[18] Kotecha N, Bruck K, Lu W, Shah N. Pathway knowledge base: An integrated pathway resource using BioPAX. Applied Ontology. 3(4); 235-245. 2008

[19] Missier P, Sahoo S, Zhao J, Goble C and Sheth A. Janus: Semantic Provenance Infrastructure for Taverna. In *Proceedings of The third International Provenance and Annotation Workshop*, Troy, NY, U.S.A, 2010. In press

[20] Altintas I, Anand M, et al.. Understanding Collaborative Studies Through Interoperable Workflow Provenance. In *Proceedings of The third International Provenance and Annotation Workshop*, Troy, NY, U.S.A, 2010. In press

[21] McCusker J. and McGuinness D. Explorations into the Provenance of High Throughput Biomedical Experiments. In *Proceedings of The third International Provenance and Annotation Workshop*, Troy, NY, U.S.A, 2010. In press