NeuroRDF: Semantic Data Integration Strategies for Modeling Neurodegenerative Diseases

Anandhi Iyappan [1,2,†], **Shweta Bagewadi** [1,2,†,*], **Matthew Page** [3], **Martin Hofmann-Apitius** [1,2], **and Philipp Senger** [1]

[1]Fraunhofer SCAI
Schloss Birlinghoven
53754 Sankt Augustin
Germany

[2]B-IT
Dahlmannstraße 2
53113 Bonn
Germany

[3]Informatics Computational Research,
UCB Pharma, 216 Bath Rd
Slough SL1 3WE
United Kingdom

{anandhi.iyappan,shweta.bagewadi}@scai.fraunhofer.de
matthew.page@ucb.com
{martin.hofmann-apitius, philipp.senger}@scai.fraunhofer.de

Abstract

Neurodegenerative diseases are incurable and debilitating conditions with huge social and economical impact, where much is still to be learnt about the underlying molecular mechanisms. Mechanistic disease models could offer a knowledge framework to help decipher the complex interactions that occur at molecular and cellular levels. This motivates the need for development of a framework consisting of heterogeneous data coupled into different regulatory layers. Thus, enabling deeper mechanistic and medical insight into such complex diseases. Here, we describe a methodology to generate semantic web-based mechanistic disease models that allow formalization of complex research questions in order to gain disease understanding.

Data for disease model construction was integrated from publicly available (semi-) structured and unstructured data resources into a single semantic web framework called *NeuroRDF*. Different data types were considered ranging from protein-protein interactions, miRNA-target interactions, pathways, to microarrays. Furthermore, we discuss in detail the data preprocessing effort incurred, and RDF schemas implemented for building *NeuroRDF*. We illustrate the effectiveness of this approach through a real world biomedical query for biomarker identification in the context of Alzheimer′s disease (AD). Furthermore, we report on the effort and challenges faced during generation of such an indication-specific knowledge base comprising curated and quality-controlled data.

Availability: The developed RDF schemas, used ontologies, and supplementary files are available at ***http://www.scai.fraunhofer.de/neuroRDF.html***. The data generated for NeuroRDF will be publicly available through the Aetionomy project[1].

1 Introduction

Although research in neurodegenerative diseases has taken tremendous strides, the characteristic pathophysiology of chronic, irreversible neuronal cell damage has limited the treatment possibilities. However, reliable biological markers of disease and disease progression could assist in early diagnosis and treatment catered to the patient (Rachakonda *et al.*, 2004). On the contrary, low-resolution outcomes of the biomarkers due to limited availability of biological samples and lack of disease specificity has hindered the progress (Rosén *et al.*, 2013).

In silico disease models have become popular in complex disease research as they provide a framework to decipher biological responses through a holistic view. These models are capable of recapitulating prime biological properties for a given condition. Furthermore, they can function as knowledge-derived decision support systems for clinical studies, reducing the cost and risk (Rodriguez-Esteban and Loging, 2013). Ideally, a harmonized aggregation of heterogeneous data sources in the form of a model facilitates data interpretation over a large

[†]These authors contributed equally.
[*]Corresponding author.

[1]http://www.aetionomy.eu

knowledge space, but data integration is far from trivial due to increasing complexity and storage limitations (Schneider and Jimenez, 2012).

There is not one but many widely used strategies aiming at a comprehensive view of the integrated data such as data warehousing, federated databases, and web-based services (*e.g.* tranSMART (Szalma *et al.*, 2010)). Data warehousing requires human intervention and it is limited to hardware/software resources allocated and increasing maintenance cost. On the other hand, federated databases are highly dependent on the service availability and internet speed. Although, web-based services facilitates integration at large-scale with good search quality, it needs to cope with variability, incompleteness, and heterogeneity in language and formats with each of the source data repositories.

Semantic web technologies have overcome the above-described challenges up to an extent by revolutionizing the lossless exchange of data and formalizing data format (Samwald *et al.*, 2011), calling it "smart data" (Kinjo *et al.*, 2012). It is built on the W3C proposed Resource Description Framework (RDF) and XML (Extensible Markup Language), the former being a standard model. Although it faces a similar challenge as of web data, usage of automated semantic reasoners has largely been beneficial.

2 Related Work

Generating a comprehensive repository for several life science data resources (as linked data) could ease cross-data querying for life science knowledge discovery. Bio2RDF (Belleau *et al.*, 2008) is one such database platform that allows integration of different data into a common knowledge space. It uses semantic web technologies by normalizing all data entities to URIs and applying a common ontology. Linking Open Drug Data (LODD) (Samwald *et al.*, 2011) is another initiative linking drug data information from DrugBank[2] and clinical trials resources. Chem2Bio2RDF (Chen *et al.*, 2010) demonstrates the potential usage of the above two mentioned RDF repositories in the field of chemoinformatics.

Among the early users of RDF in elucidating disease pathophysiology, Shin *et al.* (2012) demonstrated how RDF can enable systematic querying of linked heterogenous data to identify common genes that are differentially regulated by volatile organic compounds in diseases, pathways, and in a specific group of patients. Qu *et al.* (2009) integrated a set of nine prior knowledge sources to devise a knowledge framework, enabling mechanistic linkage for drug re-purposing in Systemic Lupus Erythematosus (SLE).

To our knowledge there has been very limited research to use RDF for neurodegenerative diseases. Lam *et al.* (2006) made the first attempt to develop an e-Neuroscience data integration framework, AlzPharm (Lam *et al.*, 2007). They extracted AD-related drug information from BrainPharm[3] to be further integrated with manually inferred hypotheses from the scientific literature and published articles (SWAN[4]). Using such a model they demonstrated its usage in clustering AD drugs based on their molecular targets and to filter publications (claims and hypotheses) specific to Donepezil effect on treatment of AD.

Our study aims at harnessing the potential of RDF as a framework for modeling neurodegenerative diseases by enabling a close, biologically sensitive integration of multiple data types. The proposed approach, called *NeuroRDF*, shows that RDF technologies are efficient in traversing different knowledge graphs (derived from distinct resources) in an integrative manner, in order to understand the underlying disease mechanisms better. Such a formalised knowledge representation will aid the mechanistic elaboration of well discussed hypotheses in neurodegeneration. Moreover, this framework can easily be extended to other disease domains.

We believe that to obtain more precise and meaningful results from huge existing data resources, data quality is of paramount importance. Considerable effort is required to process and manually curate huge amounts of data that is required to build such a knowledge base. Therefore, we propose to construct the knowledge base specific to a disease than to a domain. Along these lines, our approach focuses to build *NeuroRDF* for Alzheimer's disease indication.

[2] http://www.drugbank.ca/

[3] http://senselab.med.yale.edu/BrainPharm

[4] http://www.w3.org/TR/hcls-swan/

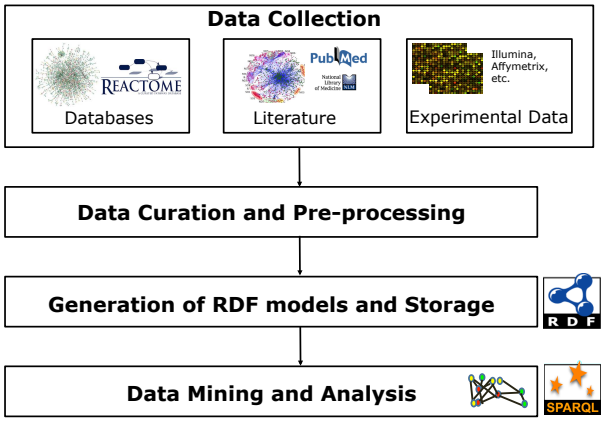

Figure 1: Schematic representation of the overall-workflow used for building *NeuroRDF*

The subsequent sections will discuss the strategies we applied to integrate AD-specific literature, databases, and gene expression information using RDF (see section 3). We will discuss in section 4 the results, and challenges faced while extracting and harmonizing the resources to build the RDF models. Further, we will demonstrate in section 5 a use case scenario on how to query different knowledge graphs to provide more meaningful biological insights into the well established hypotheses of AD disease mechanisms.

3 Methodology

The developed generic semantic web based workflow integrating heterogeneous data resources is outlined in Figure 1. This multi-layered model integrates data from various public resources such as databases, literature, and gene expression information. Harmonization of heterogenous data to build RDF models was acheived by using several data/file parsers. The workflow also includes a pre-processing step to monitor the quality of each incoming data type for specificity. The following subsections elaborate each step of the workflow and section 4 describes the details of the generated data.

3.1 Data Collection and Resources

This subsection depicts briefly different data resources integrated into *NeuroRDF*.

Database Derived Knowledge We extracted a subset of the experimentally confirmed protein-protein interactions (PPIs), assembled from 21 databases, annotated with human brain regions[18], to generate a brain PPI network representing a normal physiological state (Bossi and Lehner, 2009) (Younesi and Hofmann-Apitius, 2013). The cross-talks between molecular networks could give a hint on cascade of events participating in a disease aetiology. To uncover these associations between causal entities, we retrieved pathway information using the Reactome API[5].

Literature Derived Knowledge Data The bridging factor between researchers and scientific accomplishments are published texts, warehoused in large repositories like PubMed[6]. In order to harvest disease-specific knowledge, we used the named entity recognition (NER) system ProMiner (Fluck *et al.*, 2007) and SCAIView[7] to retrieve AD specific articles. These articles were further annotated for genes/proteins, miRNAs, and relation entity mentions. The system has been optimized for recall to reduce the false negative rate. By applying tri-occurrence based approaches, we extracted miRNA-target gene interactions (MTIs) and by using state-of-the-art machine learning based relation extraction (RE) system, we captured protein-protein interactions (Bobić *et al.*, 2012). The obtained relations have been manually filtered for false positives. The retained PPIs were additionally tagged with brain-region and experimental validation information, if available.

Experimentally Validated Knowledge Microarray-based studies provide massive potential to measure gene expression under different experimental conditions. We downloaded AD specific human datasets from GEO[8] using a simple keyword search ("alzheimer"). These datasets have been manually curated for relevant meta-data information such as age, phenotype, and tissue. We identified differentially expressed genes through SAM (Tusher *et al.*, 2001) in *R*[9] on raw data. The top 1000 genes of each dataset, ranked based on their *P-value*, were used for RDF model generation.

[5]http://www.reactome.org/
[6]http://www.ncbi.nlm.nih.gov/pubmed
[7]http://www.scaiview.com
[8]http://www.ncbi.nlm.nih.gov/geo/
[9]http://www.r-project.org

3.2 Data Curation and Pre-processing

Considering that recall-optimized automated NER and RE systems were used, the results are prone to a high false positive rate (FPs) with low precision (Czarnecki and Shepherd, 2014). Especially when considering the full text articles, FPs frequency is higher. It is not straightforward to use these systems for retrieval of context-specific triples such as for cell type specificity, disease state, or an indication. Furthermore, to retain healthy state PPIs that represent healthy state for humans also involved human effort.

Applying an automated approach to extract the meta-data annotations for microarray data such as age, phenotype, tissue, etc. is limited due to incomplete annotations and scattered information. Brazma (2009) reported that not all the data submitted to GEO or ArrayExpress[10] are MIAME compliant. Although, the published research articles are rich in annotations, a large number of experiments have missing citations (Piwowar and Chapman, 2010), which have to be be added in a manual process. The guidelines for this curation and numbers about the human effort for building such a gene expression knowledge base will be discussed in a separate publication.

Thus, to overcome above mentioned challenges all the data resources are subjected to manual curation to guarantee the interoperability and the indication specificity.

3.3 Generation of RDF Model

3.3.1 RDF Data Model

RDF allows the generation of models for processed data that exchanges information on the Web (Klapsing *et al.*, 2001). The "RDF data model" stores all the relationships between different entities as triples (subject-predicate-object). In RDF terminology, the subject, predicate, and the object are known as resources and represented by an unique "Uniform Resource Identifiers (URIs)", supporting global data exchange. Literals are the values mapped to these resources. Ontologies, similar to controlled vocabularies, explicitly describe the terms with a formal meaning to link the resources.

[10]http://www.ebi.ac.uk/arrayexpress/

3.3.2 RDF Schemas

We constructed the RDF schemas[18] by abiding the standard RDF graph notation where *ellipse* represents *Resource*, an *arrow* for *Property*, and *rectangle* for *Literal* (see Figure 2 for details). In all the RDF schemas, we have maintained a common resource representation for the "Gene", namespace adapted from OpenPhacts[11]. For the namespaces with no available ontologies, we created a namespace, called "SCAI".

For diseased PPI and MTI models, apart from the details of interactions, we additionally included evidences from articles and experiments that support these interactions. Similarly, for healthy PPI schema we included brain region and literature evidences. For microarray data, meta-annotation details have been attached to the sample. For better reasoning, quantitative values retrieved from statistical analysis are linked to genes, *cf.* Figure 2[18]. The RDF schema encoding the pathway information represents the relations between a gene and all related pathways.

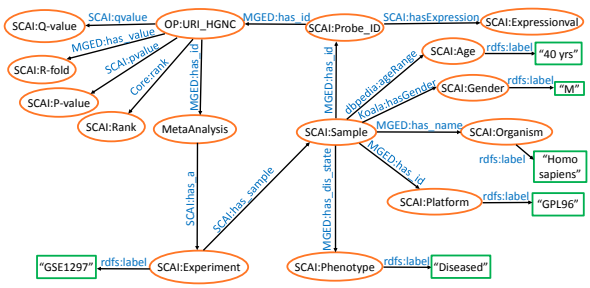

Figure 2: RDF schema for microarray data representation (*ellipse* represents *Resource*, an *arrow* for *Property*, and *rectangle* for *Literal*).

3.3.3 Construction, Validation and Storage of RDF Models

We modeled all the generated schemas in Java using the Apache Jena API[12]. *Resources*, and *Properties* are created using the corresponding in-built methods in the API and with the help of Schemagen[13]. All ontologies used in our models were converted into Java classes.

In order to check for the correctness of our generated RDF models, we made use of the online service

[11]http://www.openphacts.org/

[12]https://jena.apache.org/

[13]http://jena.apache.org/documentation/tools/schemagen.html

RDF validator[14]. Using such a service, we verified the models using both graph and triples representation.

Triple stores, such as Virtuoso[15], provides an opportunity to store individual or integrated RDF models. Taking advantage of this, we stored all our RDF models as individual graphs in a single Virtuoso instance. Using the common URI (e.g. "Gene" identifier) as the connecting link between these models, it is possible to traverse through them integratively.

3.4 Data Mining and Analysis

In RDF, all the stored triples are accessible using a common query language, SPARQL (SPARQL Protocol and RDF Query Language)[16]. We generated a Java library with embedded SPARQL queries to ask our endpoint biologically relevant questions. Individual model queries have been integrated as nested queries into one bigger query. Each query uses the common Gene URI namespace to pass on the results used to the next nested query. One possibilty to visualize the query results is the SemScape[17] Cytoscape plugin showing the return values as (sub-)graphs again.

4 Results

4.1 Data Collection

A single virtuoso endpoint stores all constructed RDF models, whose details are described in the following subsections. The healthy PPI RDF model consists of 7255 genes and 45284 interactions occurring in 15 brain regions. Out of the 45284 PPI interactions, about 36000 interactions are found to occur in specific brain regions.

The Alzheimer's disease PPI RDF network consists of 303 genes and 339 interactions along with 167 in-vitro and two in-vivo literature evidences.

In the literature derived miRNA RDF network, there are 29 miRNAs participating in 111 relations with 28 genes.

The generated microarray RDF model consists of one manually selected GEO experiment (GSE1297), for incipient AD as an example. In total there are 31 samples, 22 disease and nine normal. Among these, 19 samples belong to female and 12 samples to male donors.

For all the unique genes present in the endpoint, we derive an additional RDF network for pathway data from Reactome. Thus, for 151 genes we retrieve 280 unique pathways (with genes participating in more than one pathway).

4.2 Data Curation

Although, we used state-of-the-art relation extraction system extracted PPIs and MTIs needs manual filtering for context specificity, *i.e.* occurring in AD. We observed that only about 10-30% of the extracted PPIs and MTIs are truly relevant to AD over several iterations.

For retaining healthy state PPIs from databases, manual inspection for the specificity to healthy condition tested in certain pre-listed experiments needed 10-15 months of human effort.

Similarly, for microarray data, we invested effort to automate meta-data information extraction. However, the meta-data information is scattered in GEO website, publication, supplementary material, figures, etc. As a consequence, manual effort of about 30 minutes to 2 hours per experiment (depending on the availability and number of samples) was needed to retrieve the relevant information.

4.3 RDF Models

Table 1 summarizes the content of the generated Triple Store by providing some statistics of all integrated networks. Uploading and querying these models were not computationally expensive due to low set of relations and relatively small file size.

5 Use Case: Discovery of Integrated Data within Alzheimer's Disease

Here we report an application demonstrating the usage of the integrated data repository for facilitating data mining and analysis. This example shows the advantages of using RDF by intuitively querying all integrated data layers, connected by common namespaces.

A commonly discussed pathophysiological mechanism of AD concerns the faulty processing of amyloid precursor protein (APP) forming Aβ plaques,

[14] http://www.w3.org/RDF/Validator/
[15] http://virtuoso.openlinksw.com/
[16] http://sparql.org/
[17] http://apps.cytoscape.org/apps/semscape

Models	No. of Triples	No. of Entities	No. of Properties	Size (in MB)
Alzheimer's Disease PPI	8353	19900	11	0.894
Healthy State PPI	1204194	78852	11	99.102
MTI	667	300	5	0.095
Microarray	609991	155036	16	532.129
Pathway	1069	291	3	0.123

Table 1: Statistics of Virtuoso endpoint containing generated RDF models

leading to neurotoxicity. Over the last 25 years, molecules designed to target defective amyloid biology have not been successful in late-phase drug trials (Golde *et al.*, 2011). Thus, calling for a deeper understanding of the in-direct Aβ accumulation mechanism. A set of filters, expressed as biological conditions, have been defined to focus on the genes underlying APP related mechanisms through convergence of different data layers:

- Common to healthy and diseased state so as to derive mechanistic interpretations significant to both states
- Targeted by at least one miRNA, due to its implication in post-transcription modulation
- Up or down-regulated in human gene expression studies
- At least one of its first neighbor being differentially expressed in specific brain region, to determine local neighborhood cohesiveness
- Participating in pathways that are linked to neurodegeneration

Box 1 is an example SPARQL query syntax used to obtain common genes between healthy and AD-PPI networks. Similar querying has been applied to build a system of faceted searches to combine the above-defined conditions. The query in Box 1 resulted in 230 intersecting genes and by further filtering for genes targeted by miRNAs, we obtained 13 genes. For each of these 13 genes, we queried the microarray data to extract the fold change values. Only APP and LRP1 genes were dysregulated in the selected microarray data. APP, BACE1, MAPK3, LRP1, and CASP3 directly interact with genes that are differentially expressed. The details of the quantitative values calculated are given in Table 2. The retrieved pathways[18] show involvement

```
SELECT COUNT (DISTINCT(?Gene3)) ?Gene2 WHERE {
   GRAPH    <http://localhost:8890/Healthy_PPI> {
      ?Gene <http://purl.uniprot.org/core/encodedBy> ?p } .
   GRAPH <http://localhost:8890/Diseased_PPI> {
      ?Gene2 <http://purl.uniprot.org/core/encodedBy> ?p1 } .
   GRAPH <http://localhost:8890/Diseased_PPI> {
      ?Int <http://purl.uniprot.org/core/participant> ?p1  .
      ?Int <http://purl.uniprot.org/core/participant> ?p2  .
      ?Gene3 <http://purl.uniprot.org/core/encodedBy> ?p2 .
   FILTER (!(?p1=?p2))
   }
      FILTER (?Gene=?Gene2)
}
```

Box 1: SPARQL query syntax to retrieve intersecting genes between healthy and AD network

of these genes in neurodegeneration. Further, we created a sub-network for these genes listed in Table 2.

Interestingly, we identified two sub-networks that provide new direction for investigation in AD pathology (*cf.* Figure 3). Inspecting for literature evidences[18], we observed one publication for the first trio (PRNP-APP-BACE1) and none for the second trio (LRP1-APP-SHC1) in AD within human context (*cf.* Figure 3). However, there are additional evidences in other species and cell lines for these two trios. It is reported that binding of tyrosine phosphorylated LRP1 to SHC1 transfers the SHC1 to the plasma membrane for undergoing phosphorylation and thereby activating Ras, which is a possible cause of AD onset (Woldt *et al.*, 2011). On the other hand, it is shown that the PRNP gene modulates the APP through post-translational modification and inhibition of the BACE1, possibly causing amyloid- neurotoxicity (Lewis *et al.*, 2012).

Thus, showing the advantage of such a proposed

[18]Detailed information available as supplementary file.

Intersecting Gene Symbols	#miRNAs Targeting	Fold Change	Differentially Expressed First Neighbors
APP	18	1.13	MMP2, SHC1, PRNP, NUCB1, LRP1, MMP14, PLD1, NCL
BACE1	9	-	PRNP, APP
MAPK3	1	-	STAT1
LRP1	1	1.32	SHC1, APP
CASP3	1	-	TARDBP, APP

Table 2: Statistics of final list of genes derived from the SPARQL queries

methodology for deriving new knowledge from existing data. As a result, we can speculate new mechanisms involved in complex diseases.

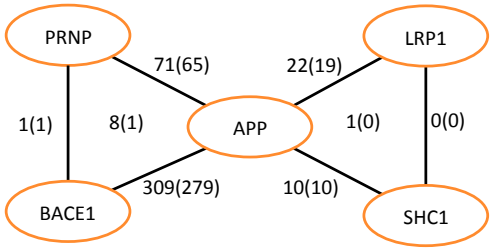

Figure 3: Extracted subnetworks with literature statistics (numbers represent count of articles supporting the interaction, in brackets are count of articles specific to human)

6 Conclusion and Future Work

The proposed integrative approach takes advantage of the well-known and highly accepted RDF technology to integrate data from various sources within a specific indication area. From our perspective it is necessary to focus on one indication or at least a group of indications to build such a knowledge base for precise modeling and analysis due to the high curation effort one has to spend in order to reach the necessary detail level. We showed how to harmonize three major heterogeneous resources (databases, gene expression, and literature) used in the research area to generate hypotheses for underlying disease mechanisms. This approach supports the tackling of ever-growing data without compromising over quality. Furthermore, new data resources can be included without altering the overall framework. The usage of well-accepted ontologies provide the advantage for further integration of external resources and databases (e.g. federated queries).

However, we are aware of several limitations of the presented method itself but also of the data used. One open point is the amendment of the RDF schema in order to improve the interoperability of the models and to expand the provided information. Furthermore, we would like to (semi-)automatically update the knowledge base and expand the networks with new incoming data. Additionally, disease modeling specific ontologies are required for such a framework. Another future goal is to work on the quality and the completeness of the underlying data resources. For example, we want to harvest full text for MTIs and robust integration of the all relevant (irrespective of the platforms) microarray experiments with stage-specific information. Inclusion of next generation sequencing, proteomics, and SNP data could enhance the quality and specificity of derived hypotheses. The presented methodology, if further optimized, is also capable of automating computer-aided translation of imaging information for personalized diagnosis and prognosis.

We would like to conclude that integrative approaches using RDF are powerful techniques to locate biologically meaningful sub-networks from highly heterogeneous and voluminous data within a well-defined indication area.

Acknowledgment

We are thankful to Erfan Younesi and Ashutosh Malhotra for providing the Healthy State PPI and AD-PPI network respectively for this work. We also want to thank Christian Ebeling for his support building the resources for microarray data.

References

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, **41**(5), 706–16.

Bobić, T., Klinger, R., Thomas, P., and Hofmann-Apitius, M. (2012). Improving distantly supervised extraction of drug-drug and protein-protein interactions. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 35–43, Avignon, France. Association for Computational Linguistics.

Bossi, A. and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular systems biology*, **5**(260), 260.

Brazma, A. (2009). Minimum Information About a Microarray Experiment (MIAME)–successes, failures, challenges. *TheScientificWorldJournal*, **9**, 420–3.

Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., and Wild, D. J. (2010). Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC bioinformatics*, **11**, 255.

Czarnecki, J. and Shepherd, A. (2014). Mining biological networks from full-text articles. In V. D. Kumar and H. J. Tipney, editors, *Biomedical Literature Mining*, volume 1159 of *Methods in Molecular Biology*, pages 135–145. Springer New York.

Fluck, J., Mevissen, H. T., Oster, M., and Hofmann-Apitius, M. (2007). ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 149–151, Madrid, Spain.

Golde, T. E., Schneider, L. S., and Koo, E. H. (2011). Anti-Aβ therapeutics in Alzheimers disease: The Need for a Paradigm Shift. *Neuron*, **69**(2), 203–213.

Kinjo, A. R., Suzuki, H., Yamashita, R., Ikegawa, Y., Kudou, T., Igarashi, R., Kengaku, Y., Cho, H., Standley, D. M., Nakagawa, A., and Nakamura, H. (2012). Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Research*, **40**(Database issue), D453–60.

Klapsing, R., Neumann, G., and Conen, W. (2001). Semantics in web engineering: Applying the resource description framework. *IEEE MultiMedia*, **8**(2), 62–68.

Lam, H. Y. K., Marenco, L., Clark, T., Gao, Y., Kinoshita, J., Shepherd, G., Miller, P., Wu, E., Wong, G., and Liu, N. (2006). Semantic Web Meets e-Neuroscience : An RDF Use Case. In *Semantic Web - ASWC 2006: First Asian Semantic Web Conference*, pages 158–170.

Lam, H. Y. K., Marenco, L., Clark, T., Gao, Y., Kinoshita, J., Shepherd, G., Miller, P., Wu, E., Wong, G. T., Liu, N., Crasto, C., Morse, T., Stephens, S., and Cheung, K.-H. (2007). AlzPharm: integration of neurodegeneration data using RDF. *BMC Bioinformatics*, **8 Suppl 3**, S4.

Lewis, V., Whitehouse, I. J., Baybutt, H., Manson, J. C., Collins, S. J., and Hooper, N. M. (2012). Cellular prion protein expression is not regulated by the alzheimer's amyloid precursor protein intracellular domain. *PLoS ONE*, **7**(2), e31754.

Piwowar, H. and Chapman, W. (2010). Recall and bias of retrieving gene expression microarray datasets through pubmed identifiers. *Journal of Biomedical Discovery and Collaboration*, **5**(0).

Qu, X. a., Gudivada, R. C., Jegga, A. G., Neumann, E. K., and Aronow, B. J. (2009). Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics*, **10 Suppl 5**, S4.

Rachakonda, V., Pan, T. H., and Le, W. D. (2004). Biomarkers of neurodegenerative disorders: how good are they? *Cell Research*, **14**(5), 347–58.

Rodriguez-Esteban, R. and Loging, W. T. (2013). Quantifying the complexity of medical research. *Bioinformatics (Oxford, England)*, **29**(22), 2918–24.

Rosén, C., Hansson, O., Blennow, K., and Zetterberg, H. (2013). Fluid biomarkers in Alzheimer's disease - current concepts. *Molecular Neurodegeneration*, **8**, 20.

Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C. S., Willighagen, E., Hajagos, J., Marshall, M. S., Prud'hommeaux, E., Hassenzadeh, O., Pichler, E., and Stephens, S. (2011). Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*, **3**(1), 19.

Schneider, M. V. and Jimenez, R. C. (2012). Teaching the fundamentals of biological data integration using classroom games. *PLoS Computational Biology*, **8**(12), e1002789.

Shin, G.-H., Kang, Y.-K., Lee, S.-H., Kim, S. J., Hwang, S. Y., Nam, S.-W., Ryu, J.-C., and Kang, B.-C. (2012). mRNA-centric semantic modeling for finding molecular signature of trace chemical in human blood. *Molecular & Cellular Toxicology*, **8**(1), 35–41.

Szalma, S., Koka, V., Khasanova, T., and Perakslis, E. D. (2010). Effective knowledge management in translational medicine. *Journal of Translational Medicine*, **8**, 68.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(9), 5116–21.

Woldt, E., Matz, R. L., Terrand, J., Mlih, M., Gracia, C., Foppolo, S., Martin, S., Bruban, V., Ji, J., Velot, E., Herz, J., and Boucher, P. (2011). Differential signaling by adaptor molecules LRP1 and ShcA regulates adipogenesis by the insulin-like growth factor-1 receptor. *J Biol Chem*, **286**(19), 16775–82.

Younesi, E. and Hofmann-Apitius, M. (2013). Biomarker-guided translation of brain imaging into disease pathway models. *Sci. Rep.*, **3**.