Leveraging SADI Semantic Web Services to exploit fish ecotoxicology data

Matthew M. Hindle[1], Alexandre Riazanov[1], Edward S. Goudreau[2], Christopher J. Martyniuk[2], Christopher J. O. Baker[1]

[1] Department of Computer Science & Applied Statistics
[2] Canadian Rivers Institute and Department of Biology
University of New Brunswick, Saint John, New Brunswick, E2L 4L5, Canada
{hindlem, alexr, t969c, cmartyn, bakerc}@unb.ca

Abstract. In order to interpret experimental Omics-data, ecotoxicologists are faced with an array of disconnected bioinformatics databases and algorithms. These include tools for microarray analysis, gene annotation, functional gene set enrichment, and network analysis. Drawing together these Web tools and resources is a frequently labour-consuming technical exercise in identifying links across database records and the connecting input and output formats of tools. Interpreting experimental Omics-data in the context of the current available knowledge and methodologies from a single query platform with explicit semantics would be a valuable asset for toxicology in the analysis of DNA, transcriptomics, proteomic, and metabolomic experimental data.

Methods: We have created 30+ SADI semantic Web Services, resources and tools pertinent to the interpretation of Omics toxicological data. These services encompass a wide range of algorithms, domains and databases, including sequence alignment and protein domain finding tools (e.g. BLAST, HMMR3, and InterProScan), databases containing experimentally validated protein functions (e.g. ZFIN and MGI), and central repositories of sequence and microarray data (e.g. ArrayExpress and NCBI-RefSeq). All these services can be leveraged through SPARQL queries submitted to the SHARE query engine. This paradigm provides a single access-point on the Web for a toxicologist to submit semantically rich queries that are resolved using the relevant databases and tools. This frees the ecotoxicologist from learning unnecessary details concerning tool interfaces and the semantic idiosyncrasies of databases.

Results: We present a series of example queries for ecotoxicology, which facilitate the interpretation of transcriptomics data in the context of public knowledge and current tools. These queries include common tasks, specific to a user's experimental data set, such as gene ontology annotation of probes on a custom microarray experiment for an aquatic species of interest.

Keywords: SADI, SHARE, Semantic Web Services, Ecotoxicology, Fish, Toxicology

1 Introduction

Toxicology is increasingly a systems discipline, requiring the analysis of multi-scale Omics data [2, 17, 25]. This typically require tools and databases that can be leveraged for tasks such as microarray analysis, gene annotation, functional gene set enrichment, and network analysis. However, in order to meet these requirements toxicologists are faced with a bewildering array of disconnected bioinformatics resources. Drawing together and mastering these Web tools and resources is frequently an unnecessary and frustrating technical exercise in identifying common links across database records and the connecting input and output formats of bioinformatics tools. Interpreting experimental Omics-data in the context of the current available knowledge and methodologies from a single query platform with explicit semantics would be an invaluable asset to ecotoxicologists in the analysis of their DNA, transcriptomics, proteomic, and metabolomic experimental data. Working towards such a unified semantic framework in ecotoxicology, would free toxicologists from wasting time on technical and semantic idiosyncrasies, and enable the environmental toxicologist to synthesize Omics information to better predict risks associated with chemical exposures.

There are many existing approaches to integrating biological data sets. Project like Bio2RDF [4] and Linked Life Data [30] have used semantic technologies to build mash-ups of current biological information. However, many biological application cases also require the integration of bioinformatics tools and algorithms. Semantic Web Service architectures [9, 12, 16, 26] are an elegant solution to exposing both knowledge and algorithms in a semantically explicit framework. Services can be leveraged as components in complex bioinformatics analysis pipelines. This paper presents an initial framework of SADI Web Services for the ecotoxicology domain and example queries which demonstrate how such a framework can be leveraged to retrieve relevant information. Together with existing SADI use-cases [3, 7, 21], these example queries demonstrate the utility of SADI semantic Web Services in solving the problems of resource and tool fragmentation, and semantic heterogeneity in the life sciences.

1.1 What are SADI Web Services?

Most conventional Web Services produce an output without making an explicit semantic connection to the input data. Web Services built using the SADI framework [26] make the semantics of this relationship explicit. SADI is a set of conventions for creating Semantic Web Services, which as a consequence of their explicit semantics, can be automatically discovered and orchestrated. An RDF graph forms the service input and has some URI node designated as a central node. The whole input RDF graph is considered a description of this central node. Exactly the same node is always present in the output RDF graph and becomes the central output node. The sole function of a SADI service is therefore to decorate this central input node with new properties, which are asserted in the output RDF graph.

The classes and properties which a SADI Web Service accepts as input and computes as output must reference a defined input and output class in an OWL [31] ontology. The inputs and outputs of the Web Service are therefore always clearly defined and the behavior of the Web Service is formally specified. For example, Fig. 1 shows a SADI Web Service and an example of the RDF input and output accepted by the service. The input and output classes are defined in the service ontology.

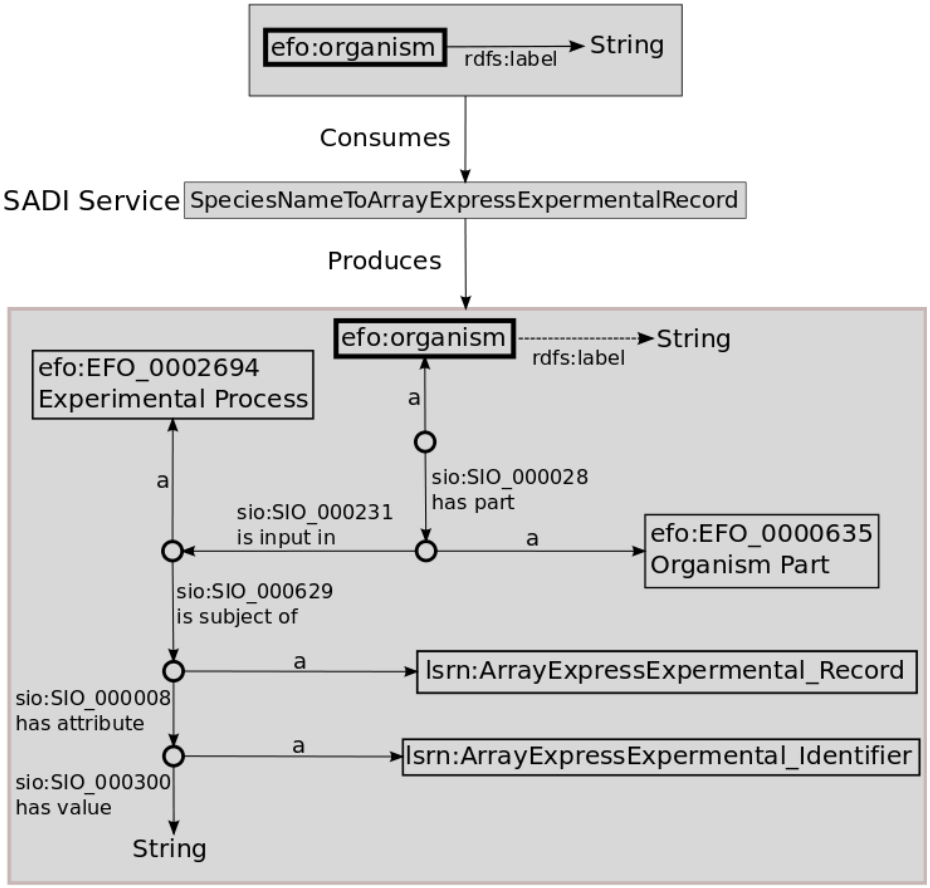


Fig. 1 A SADI service which uses input and output classes from the service ontology: http://unbsj.biordf.net/fishtox/arrayexpress-sadi-service-ontology.owl. The service consumes an OWL class, which must be a subclass of efo:organism and have an attach label. The modeling of the input is defined in ExperimentAnnotatedSpecies_Input. The service decorates the OWL organism class with the individuals of that species that are inputs for microarray experiments in ArrayExpress. The modeling of the output is defined by the ExperimentAnnotatedSpecies_Output class. A key to the prefixes used in this figure can be found in Table 1. The bold framed RDF class indicates the central node in the input and output graph.

The explicit nature of SADI service semantics means they can be automatically enacted by client software. In this paper we use the SHARE client [23], which computes SPARQL queries by picking and calling suitable SADI Web Services from a SADI service registry. Therefore, SPARQL queries can be written based on an understanding of the ontological primitives referred to in service semantic descriptions, available from the registry. However, browsing the services is often a useful exercise as it gives a good idea of what data is available.

1.2 The tools and databases wrapped as SADI Web Services

In order to improve predictive abilities, ecotoxicologists are becoming more interested in the pathways and associations regulated by a specific chemical (*e.g.* adverse pathways of toxicity). In order to provide a core bioinformatics toolbox for ecotoxicology, our initial SADI Web Services provide information pertinent to the analysis of exotoxicology microarrays. Specifically, we prioritize services which facilitate (1) comparison of an experimental dataset with other published transcriptomics data, and (2) sequence transcript information retrieval in the form of Pfam [11] protein domain, and Gene Ontology (GO) functional annotation. These knowledge domains provide the subject for the example queries, described in the results.

An important requirement for the analysis of a fish toxicological dataset is the ability to compare experimental results with existing published data. This has the potential to provide valuable insights into transcriptomics datasets by elucidating similarities and differences with transcriptomes that were subject to similar experimental conditions, such as the concentration of chemical or duration of exposure regime. ArrayExpress [20] is a database of functional genomics experiments which includes a large number of microarrays. It includes data on microarray platforms, as well as data recording individual experiments and their parameters. It provides Web Services that can be readily wrapped with SADI Web Services, which effectively provide a layer that adds explicit semantics.

Another requirement for our ecotoxicology use case is the annotation of microarray sequences with Pfam domains and GO functional annotation. In order to achieve this, SADI Web Services were required that exposed HMMR3 [8] and BLAST functionality. Microarray sequences are often derived from assembled EST sequences, and this is particularly true for the many custom arrays for fish. Consequently sequences may be incomplete and contain missing gene fragments, which introduce shifts in the reading frame. This makes the process of finding the correct open reading frame (ORF), which encodes the protein, challenging. It was therefore a priority to include a ORF prediction tool such as ORF-Predictor [19].

Sequence functional annotation with GO also requires the retrieval of experimentally derived annotations from model organism databases. We prioritized annotations for *Danio rerio* and *Mus musculus*, based on the evolutionary distance to fish and abundance of experimental annotation, respectively.

The Zebrafish Model Organism Database (ZFIN) [6], is the main data repository for the *Danio rerio* genome. *Danio rerio* is one of the most important model organisms for teleost fish and is used as a model for growth and development, pharmacology and toxicology studies [14, 22, 24]. ZFIN contains a repository of reference gene models, together with mappings to most of the sequence repositories. They also contribute a set of experimental and electronically inferred GO annotations for genes.

Mouse Genome Informatics (MGI) [5] is the main data repository for information concerning the *Mus musculus* genome. It contains a list of reference gene models, and external references to the main sequence repositories. It is the largest source of experimentally verified GO annotations for genes, which motivated us to included it as a data source for SADI Web Services.

2 Methods

This section describes the prior modeling and SADI Web Services, which were leveraged by the example queries in this use case. Defining an appropriately expressive model for the RDF that will be consumed and produced by services is crucial for enabling interoperability with other services, and flexible querying.

2.1 Reuse of existing upper and domain Ontologies

In order to improve the re-usability of our SADI Web Services, wherever possible we reference existing upper and domain ontologies. Table 1 lists the ontologies used by the SADI Web Services. The Semanticscience Integrated Ontology (SIO) provides a broad set of classes and properties, and is used extensively by other SADI Web Services. The Life Science Resource Name (LSRN) provides classes for defining database records and identifiers. It also uses the SIO ontology as an upper ontology. SIO and LSRN are our preferential upper ontologies for modeling services. The Experimental Factor Ontology (EFO), provides classes and properties for describing sample variables in experiments [15]. It has been used extensively for the Gene Expression Atlas [13], and in the Semantic Web Atlas Project [1]. We reuse EFO to encourage interoperability with the modeling provided by these projects. Our application ontologies mainly contain input and output class definitions. Where possible we have minimized the creation of any new classes or relations in these service ontologies.

Table. 1 Ontologies and Prefixes used in the SADI Web Services

Prefix	URL	Type
lsrn	http://purl.oclc.org/SADI/LSRN/	Upper
sio	http://semanticscience.org/resource/	Upper
efo	http://www.ebi.ac.uk/efo/	Domain
blastso	http://unbsj.biordf.net/fishtox/BLAST-sadi-service-ontology.owl#	Application
hmmrso	http://unbsj.biordf.net/fishtox/HMMR-sadi-service-ontology.owl#	Application
goaso	http://unbsj.biordf.net/fishtox/GOA-sadi-service-ontology.owl#	Application
microarrayso	http://unbsj.biordf.net/fishtox/arrayexpress-sadi-service-ontology.owl#	Application
tsso	http://unbsj.biordf.net/fishtox/record-translation-sadi-service-ontology.owl#	Application
stso	http://unbsj.biordf.net/fishtox/seq-tools-sadi-service-ontology.owl	Application

2.2 Modeling schematics

Fig 2. shows a schematic of the main classes and properties used to model RDF input and output for SADI Web Services. The schematic can be used to design SPARQL queries for the SHARE client. Some secondary classes and properties, such as BLAST and HMMR alignment scores, have been omitted. Also, the schematic does not show potential connections to classes and properties provided by other published SADI Web Services. The semantic richness of our modeling enables a greater expressiveness in writing SPARQL queries. It also reduces the need to re-model for new use cases when further SADI Web Services are added or become available.

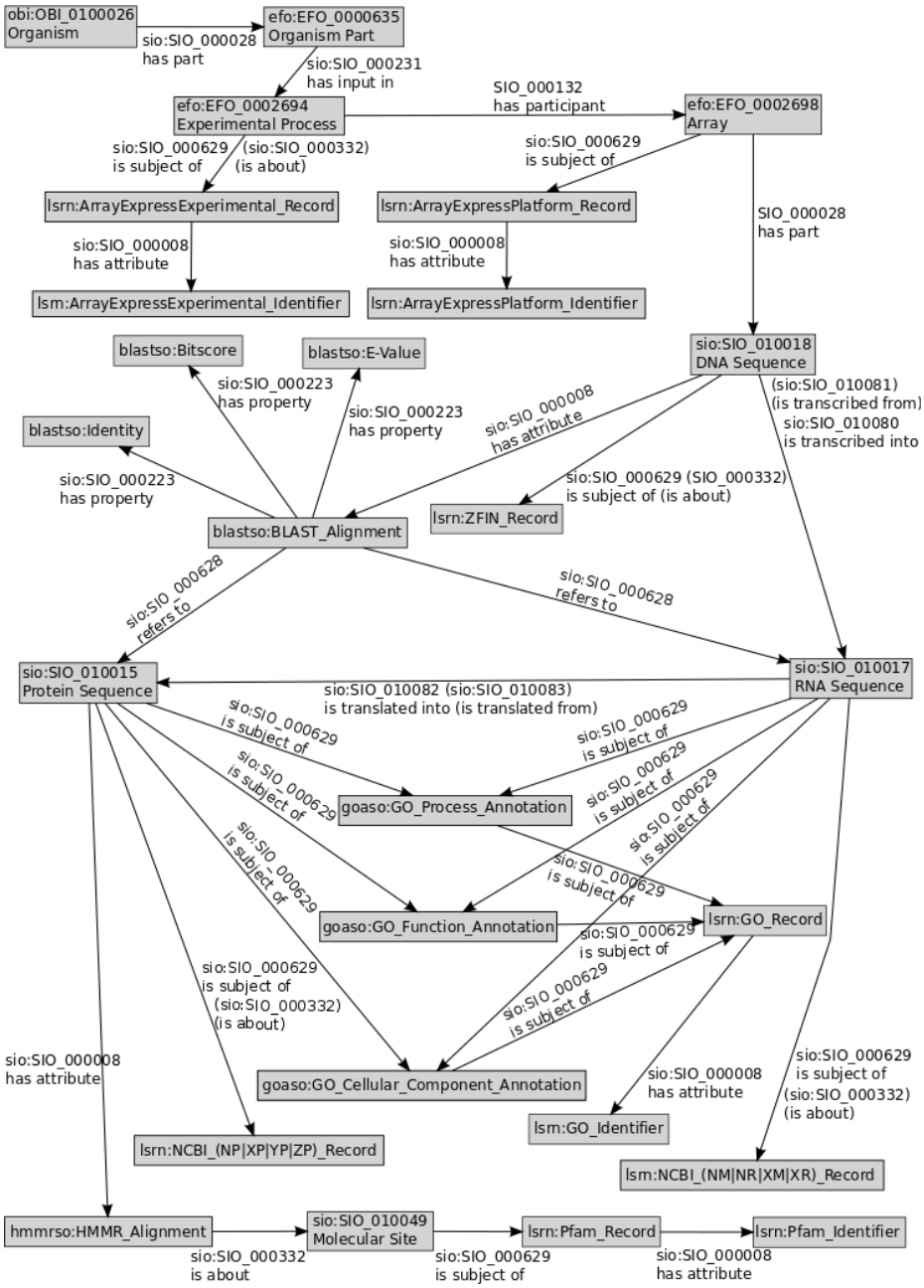


Fig 2. A schematic of the main OWL classes and properties used to expose data in the described SADI Web Services. A key to the prefixes used in this figure can be found in Table 1. The properties which are given in brackets are the inverse of the properties for the given direction, indicated by the connecting arrow.

2.3 SADI Web Services for fish research and aquatic ecotoxicology

In total we created 32 SADI Web Services which exposed information from five database: ArrayExpress, ZFIN, MGI, RefSeq, Pfam, and GO. We also exposed BLASTn, BLASTx, BLASTp, HMMR3 and ORF-Predictor tools. These services are too numerous to describe all but a selection in detail here, however a description of each is provided at http://unbsj.biordf.net/FISHTOX-SADIServices. A SHARE client has been made available to query these service at http://unbsj.biordf.net/cardioSHARE-fishtox.

Where possible we wrapped existing Web Services, provided by databases and tools, as SADI Web Services. This provides live data, which ensures results are current, and avoids the maintenance cost associated with data mirrors.

Four ArrayExpress SADI Web Services, one of which has been described already (Fig. 1), were created by wrapping the Web-Services provided by ArrayExpress. Information exposed was modeled using a combination of the existing EFO ontology (which the database supports natively) and SIO properties.

HMMR3 SADI Web Services were provided by wrapping the Web Services provided by janelia [10]. The input class of the service is a 'protein sequence' (sio:SIO_010015) and output class is defined in the hmmrso (Table 1) ontology as a class that:

```
'has attribute' min 1 (HMMR_Alignment that ('is about' min 1
('molecular site' that ('is subject of' min 1 Pfam_Record))))
```

Similarly SADI Web Services for BLAST were created by wrapping NCBI Web Services. The input to these services was either a 'protein sequence' (sio:SIO_010015) or a 'nucleic acid sequence' (sio:SIO_010016) depending on the variant of BLAST. The output is defined using an alignment class, in an approach similar to HMMR3 services. For example, the output for the BLASTx is defined in the blastso (Table 1) ontology as a class that:

```
'has attribute' some (BLAST_Alignment that ('refers to' min 1 ('protein
sequence' that ('is subject of' min 1 (NCBI_NP_Record or NCBI_AP_Record
or NCBI_XP_Record or NCBI_YP_Record or NCBI_ZP_Record)))))
```

Translation between any two database records is handled by modeling the relation between the sequences which they concern. For example, an lsrn:ZFIN_Record concerns some genomic sequence corresponding to a gene model. The SADI translation service consuming instances of this class as input, defines the relationship between the input and an NCBI protein record in RefSeq via the following output class tsso:RefSeq_Protein_Annotated_Record_Output:

```
'is about' min 1 ('deoxyribonucleic acid sequence' that ('is
transcribed into' min 1 ('ribonucleic acid sequence' that ('is
translated into' min 1 ('protein sequence' that ('is subject of' min 1
(NCBI_NP_Record or NCBI_AP_Record or NCBI_XP_Record or NCBI_YP_Record
or NCBI_ZP_Record)))))))
```

GO annotation SADI Web Services were created by directly RDFizing ZFIN and MGI annotations published on the GO website [29]. They annotate both

lsrn:ZFIN_Record and lsrn:MGI_Record classes. The definition of the output class is complex as the GO annotation can reference the function, process, or cellular compartment of the RNA or Protein product of the DNA which is the subject of ZFIN or MGI records.

3 Results

In this section we present three example queries, which address the types of questions pertinent to the analysis of gene expression data. However, the SADI Web Services we have built, and the modelling we employ, is not limited to these examples. Any number of combinations of these SADI Web Services, together with the growing number of public SADI Web Services, can be used to produce many useful queries. In this paper we focus on a few example queries based around the analysis of fish toxicology data, however these methodologies are widely applicable to gene expression analysis.

These queries are enacted by the SHARE client, which computes queries by picking and calling suitable SADI Web Services from a dedicated registry of fish toxicology-related services. The SHARE client Web interface reports results in tabular form and as a downloadable RDF graph.

3.1 Query I: Leveraging ORF finding algorithms to detect Pfam domains

After the gene sequences of interest have been identified, a common requirement is to classify these genes according to the protein domains, which they encode. This is often a non trivial task for microarray sequences, which are frequently derived from assembled EST sequences. Consequently sequences may be incomplete and contain missing gene fragments, which introduce shifts in the reading frame across the sequence. This makes the process of finding the correct open reading frame (ORF) which encodes the protein challenging. Combining the output of a ORF prediction tool, together with the HMMR3 algorithm, without scripting, would require a great deal of manual work for a biologist, which becomes insurmountable for anything but a trivial number of sequences. The following SPARQL query annotates Pfam domains for the ten most significantly regulated genes in *Micropterus salmoides*, relative to the control, under dieldrin-induced stress [18]. In order to compute the query SHARE calls three services. The first service decorates the DNA sequences on the chip with RNA. The RNA is then passed through the ORF prediction service to decorate a protein sequence, which is then passed to the HMMR3 service, which adds protein domains to the RDF model.

```
1. PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2. PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3. PREFIX sio: <http://semanticscience.org/resource/>
4. SELECT ?DNA_chip_sequence ?pfam_name
5. FROM <http://unbsj.biordf.net/fishtox/TopTenLowestPvalue-DdResponsiveGenes.rdf>
6. WHERE {
```

```
7.  ?DNA_chip_sequence sio:SIO_010080 ?RNA_sequence .
8.  # (is transcribed into)
9.  ?RNA_sequence sio:SIO_010082 ?protein_sequence .
10. # (is translated into)
11. ?protein_sequence sio:SIO_000008 ?alignment .
12. # (has attribute)
13. ?alignment sio:SIO_000332 ?molecular_site .
14. # (is about)
15. ?molecular_site sio:SIO_000629 ?pfam_record .
16. # (is subject of)
17. ?pfam_record rdfs:label ?pfam_name
18. }
```

The first predicate in the query (line 7) causes SHARE to look for services indexed by the "is transcribed into" predicate. It finds a DNA2RNA SADI service which consumes a DNA sequence class and decorates this with an RNA sequence, which is attached by the "is transcribed into" property. The second predicate (line 9) is resolved by an ORF predictor service, which consumes RNA sequences, and uses sequence alignment to RefSeq proteins (BLASTx) to predict the most likely open reading frames that code for proteins. SHARE feeds the RNA sequences outputted by the DNA2RNA service into the ORF predictor SADI service, which decorates them with protein sequences attached by a "is translated into" property. The third predicate (line 11) can be resolved by the HMMR3 service, which consumes a protein sequence and produces HMMR alignments with attached Pfam protein domains. The forth and fifth predicates (line 15 and 17) are part of the output modeling of this HMMR3 service.

The SHARE client returned the answer that the domain Ribosomal_L7Ae was found on the gene UF_Msa_AF_100231. The low coverage on genes (10%) is not surprising given the species (*Largemouth Bass*), and the conservative default settings of the HMMR3 SADI service (e-value < Gathering threshold). The service is parameterized to allow these settings to be changed, but this functionality is not yet supported in SHARE. The RDF output from this query can be found at http://unbsj.biordf.net/fishtox/QueryIOutput.rdf.

3.2 Query II: Functional annotation of sequence data

One of the most powerful tools for microarray data is GO functional annotation. However, for a non-model organism like *Micropterus salmoides*, very little experimental evidence is recorded in public repositories for GO function. It is therefore necessary to infer function based on sequence similarity with known genes in model organisms. The following SPARQL query annotates the ten genes previously described in Section 3.2. To execute the query, SHARE calls the BLASTx service to find similar proteins in the RefSeq database, looks up the equivalent ZFIN and MGI Records, and then finally retrieves experimentally evidenced GO terms for these records using the corresponding SADI services from our set. The default e-value threshold for parameterized BLAST services is $1x10^{-4}$.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sio: <http://semanticscience.org/resource/>
SELECT ?DNA_chip_sequence ?zfin_or_msi_record ?go_id
FROM <http://unbsj.biordf.net/fishtox/TopTen-lowest-pvalue-DdResponsiveGenes.rdf>
```

```
WHERE {
  ?DNA_chip_sequence sio:SIO_000008 ?alignment .
  # (has attribute)
  ?alignment sio:SIO_000628 ?sequence_hit .
  # (refers to)
  ?sequence_hit sio:SIO_000629 ?refseq_record .
  # (is subject of)
  ?refseq_record sio:SIO_000332 ?RNA_Sequence .
  # (is_about)
  ?RNA_Sequence sio:SIO_010081 ?DNA_sequence .
  # (is_transcribed_from)
  ?DNA_sequence sio:SIO_000629 ?zfin_or_msi_record .
  # (is subject of)
  ?zfin_or_msi_record sio:SIO_000332 ?DNA_sequence .
  # (is about)
  ?DNA_sequence sio:SIO_010080 ?RNA_Sequence.
  # (is transcribed into)
  ?RNA_sequence sio:SIO_010082 ?protein_sequence .
  # (is translated into)
  ?protein_sequence sio:SIO_000629 ?GO_annotation .
  # (is subject of)
  ?GO_annotation  sio:SIO_000629 [rdfs:label ?go_id]
}
```

The results from SHARE indicate the presence of ribosomal processes and functions, that were experimentally evidenced in genes with sequences similar to the ten sequences being annotated. This accords well with the Pfam domains found by Query II. The query results also identified a number of additional gene functions, which include transcription factor, enzyme binding, steroid hormone receptor, cholesterol transporter, and phospholipid binding activities. The RDF output from this query can be found at http://unbsj.biordf.net/fishtox/QueryIIOutput.rdf.

3.3 Query III: Locating relevant microarray experiments

A frequent requirement of experimentalists involved in transcriptomics is the comparison of their own work with previous published experiments. Locating microarray experiments with related experimental variables is a prerequisite for further comparative analysis. In order to answer this question an experimentalist typically would use the Web-tools provided by ArrayExpress. One such example query might be “For the hypothalamus of *Micropterus salmoides*, what gene transcripts have been measured in existing experiments”. Using ArrayExpress Web based tools alone would require multiple searches and manual inspections of many experiments each of which may use different microarray platforms. The following declarative SPARQL query expresses this question formally. Note that the RDF file http://unbsj.biordf.net/fishtox/large-mouth-bass-27706.owl specified in the FROM clause contains the OWL class of the organism of interest (*Micropterus salmoides*) to instantiate the first query line. This was created from a subset of http://purl.org/obo/owl/NCBITaxon using OntoFox [28]

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX aeso:
        <http://unbsj.biordf.net/fishtox/arrayexpress-sadi-service-ontology.owl#>
SELECT ?experiment_accession ?tissue_name ?platform_accession ?gene_id
FROM <http://unbsj.biordf.net/fishtox/large-mouth-bass-27706.owl>
```

```
WHERE {
 ?organism_class aeso:has_instance ?organism_instance .
 ?organism_instance a ncbitaxon:NCBITaxon_27706 .
 ?organism_instance sio:SIO_000028 ?organism_part .
  # (has part)
  ?organism_part sio:SIO_000231 ?experimental_process .
  # (is input in)
 ?organism_part rdfs:label ?tissue_name .
  # (has value)
  ?experimental_process sio:SIO_000629 ?experimental_record .
  # (is subject_of)
  ?experimental_record sio:SIO_000008 [sio:SIO_000300 ?experiment_accession] .
  # (has attribute, has value)
  ?experimental_record sio:SIO_000332 ?experimental_process .
  # (is about)
  ?experimental_process sio:SIO_000132 ?array .
  #(has participant)
  ?array sio:SIO_000629 ?array_platform_record .
  # (is subject of)
  ?array_platform_record sio:SIO_000008 [sio:SIO_000300 ?platform_accession] .
  # (has attribute, has value)
  ?array_platform_record sio:SIO_000332 ?array .
  # (is about)
  ?array sio:SIO_000028 [rdfs:label ?gene_id] .
  # (has part)
  FILTER regex(?tissue_name, "hypothalamus", "i")
}
```

The query was submitted to the SHARE client which resolved the answer by leveraging 5 SADI Web Services which expose relevant ArrayExpress data. No understanding of the ArrayExpress semantic idiosyncrasies or data syntax was required to formulate the query. SHARE identifies two microarray experiments which meet the requirements of this query. The RDF output from this query can be found at http://unbsj.biordf.net/fishtox/QueryIIIOutput.rdf.

4 Conclusions and further work

The aim of this fish toxicology use-case was to demonstrate how a moderate number of SADI Web Services can enable diverse and powerful queries using the SHARE client. The services described exposed information from five databases and three analytical tools in a semantically rich and explicit way. They provided a single access-point for an ecotoxicologist to query data, and a unified and semantically consistent data representation. When using this framework, a ecotoxicologist would not require understanding of the semantics and technicalities of the underlying resources, in order to construct queries across databases and tools. We acknowledge that designing SPARQL queries may be beyond the reach of many biologists. However, the graphical workflow and query tools, Taverna [27] and Sentient Knowledge Explorer [32], have active SADI plug-ins under development, which may provide a solution to this interface deficiency.

In future work we will expand the SADI Web Services provided in this use-case to leverage experimental observations of gene expression. We will also provide services for common statistical methods, such as gene set enrichment analysis. In our specific

example with largemouth bass, this will enable queries such as “Which GO functions are significantly enriched in teleost fish in response to dieldrin treatment”. We also intend to develop queries which leverage some of the many public SADI Web Services, developed outside this project.

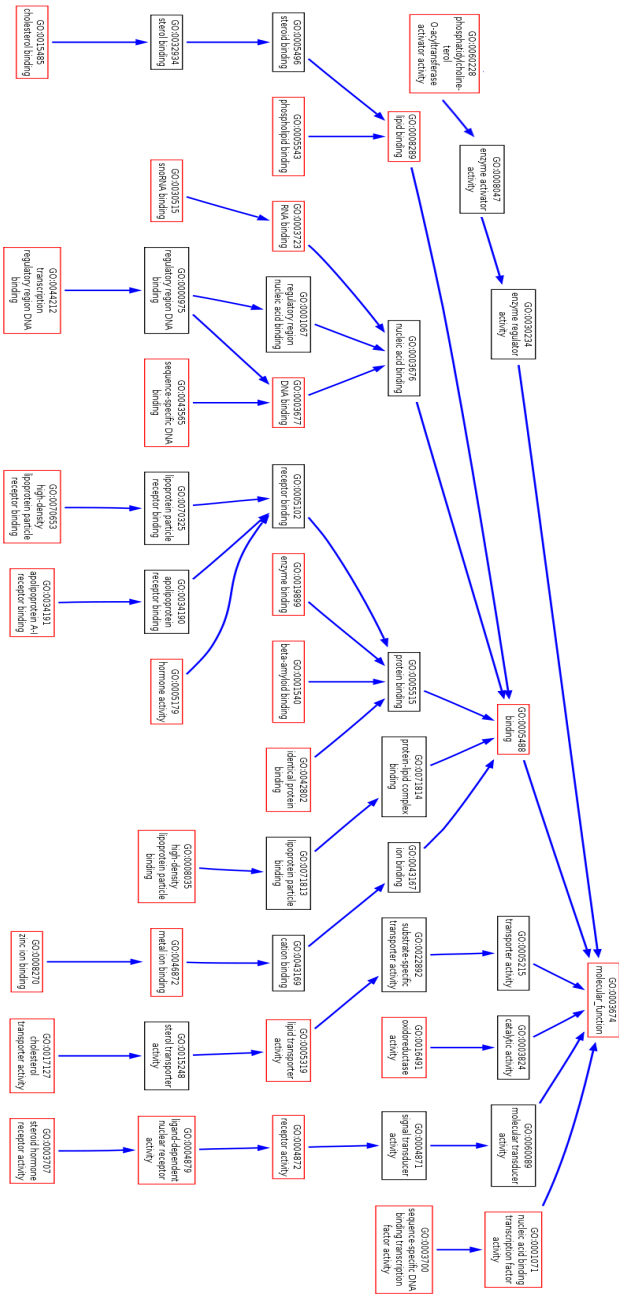
Acknowledgements

This research was funded by CANARIE NEP-2 Program (C-BRASS project). We thank Luke McCarthy and Ben Vandervalk for helping us with various SADI-related technical issues.

References

1. Adamusiak, T., Malone, J.: Semantic Web Atlas Project, http://www.ebi.ac.uk/efo/semanticweb/atlas.
2. Van Aggelen, G. et al.: Integrating Omic Technologies into Aquatic Ecological Risk Assessment and Environmental Monitoring: Hurdles, Achievements, and Future Outlook. Environ Health Perspect. (2009).
3. Baker, C.J.O.: Semantic Infrastructure for Automated Small Molecule Classification and Data Mining for Lipidomics. CSHALS. , Boston, MA (2011).
4. Belleau, F. et al.: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics. 41, 5, 706-716 (2008).
5. Blake, J.A. et al.: The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Research. 39, Database, D842-D848 (2010).
6. Bradford, Y. et al.: ZFIN: enhancements and updates to the zebrafish model organism database. Nucleic Acids Research. (2010).
7. Chepelev, L.L., Dumontier, M.: Semantic Web integration of Cheminformatics resources with the SADI framework. Journal of Cheminformatics. 3, 16 (2011).
8. Eddy, S.R.: Accelerated profile HMM searches, ftp://selab.janelia.org/pub/publications/Eddy11/Eddy11-preprint.pdf.
9. Farrell, J., Lausen, H.: Semantic Annotations for WSDL and XML Schema, http://www.w3.org/TR/sawsdl/.
10. Finn, R.D. et al.: HMMER web server: interactive sequence similarity searching. Nucleic Acids Research. (2011).
11. Finn, R.D. et al.: The Pfam protein families database. Nucleic Acids Research. 38, Database, D211-D222 (2009).
12. Gessler, D. et al.: SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services. BMC Bioinformatics. 10, 1, 309 (2009).
13. Kapushesky, M. et al.: Gene Expression Atlas at the European Bioinformatics Institute. Nucleic Acids Res. 38, Database issue, D690-D698 (2010).

14. Löhr, H., Hammerschmidt, M.: Zebrafish in endocrine systems: recent advances and implications for human disease. Annu. Rev. Physiol. 73, 183-211 (2011).
15. Malone, J. et al.: Modeling Sample Variables with an Experimental Factor Ontology. Bioinformatics. (2010).
16. Martin, D. et al.: Bringing semantics to web services: The OWL-S approach. Semantic Web Services and Web Process Composition. 26–42 (2005).
17. Martyniuk, C.J. et al.: Omics in aquatic toxicology: Not just another microarray. Environmental Toxicology and Chemistry. 30, 2, 263-264 (2011).
18. Martyniuk, C.J. et al.: Genomic and Proteomic Responses to Environmentally Relevant Exposures to Dieldrin: Indicators of Neurodegeneration? Toxicological Sciences. 117, 1, 190 (2010).
19. Min, X.J. et al.: OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic acids research. 33, suppl 2, W677 (2005).
20. Parkinson, H. et al.: ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res. 39, Database issue, D1002-D1004 (2011).
21. Riazanov, A. et al.: Deploying the Mutation Impact mining pipeline with SADI: an exploratory case study.
22. Sukardi, H. et al.: Zebrafish for drug toxicity screening: bridging the in vitro cell-based models and in vivo mammalian models. Expert Opin Drug Metab Toxicol. 7, 5, 579-589 (2011).
23. Vandervalk, B. et al.: SHARE: A Semantic Web Query Engine for Bioinformatics. The Semantic Web. 367–369 (2009).
24. Vascotto, S.G. et al.: The zebrafish's swim to fame as an experimental model in biology. Biochem. Cell Biol. 75, 5, 479-485 (1997).
25. Villeneuve, D.L., Garcia-Reyero, N.: Vision & strategy: Predictive ecotoxicology in the 21st century. Environmental Toxicology and Chemistry. 30, 1, 1-8 (2011).
26. Wilkinson, M.D. et al.: SADI Semantic Web Services-‚cause you can't always GET what you want! Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific. pp. 13–18 (2009).
27. Withers, D. et al.: Semantically-guided workflow construction in Taverna: the SADI and BioMoby plug-ins. Leveraging Applications of Formal Methods, Verification, and Validation. 301–312 (2010).
28. Xiang, Z. et al.: OntoFox: web-based support for ontology reuse. BMC Research Notes. 3, 1, 175 (2010).
29. Current Annotations, http://www.geneontology.org/GO.downloads.annotations.shtml.
30. Linked Life Data, http://linkedlifedata.com/.
31. OWL 2 Web Ontology Language Document Overview, http://www.w3.org/TR/owl2-overview/.
32. Sentient Knowledge Explorer, http://www.io-informatics.com/products/sentient-KE.html.



Appendix 1. GO Molecular functions identified by SHARE for example query III. Red bordered concepts indicate annotations found for similar sequences.