Semantic Web Atlas: Putting Gene Expression Data Into Biological Context

Simon Jupp, Helen Parkinson, and James Malone

Functional Genomics Group, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD.
jupp@ebi.ac.uk, malone@ebi.ac.uk

Abstract. We present an RDF representation of the Gene Expression Atlas (GXA) at the European Bioinformatics Institute. The RDF representation provides new opportunities for data integration, querying, error detection and curation. The GXA RDF connects to other EBI resources including Ensembl, UniProt and Reactome following linked data principles. These links enable queries across resources such as querying differential gene expression in the context of pathway and gene product functions. We show how the RDF representation of the GXA adds value to the existing data through a series of novel biological questions which interrogate the different resources and which were not previously possible through the existing GXA.

Keywords: RDF, gene protein pathway integration, Gene expression Atlas

1 Introduction

The Gene Expression Atlas database (GXA) [1] [2] at the European Bioinformatics Institute contains meta-analyses of gene expression experiments from over 3,000 publicly available biomedical investigations taken from the ArrayExpress (AE) archive [3]. Data stored in the GXA is manually curated by expert biologists and makes use of ontologies and controlled vocabularies to perform meaningful annotation across experiments. Genes are annotated using ENSEMBL [4] and experimental variables are annotated using the Experimental Factor Ontology (EFO) [5]. These annotations are then stored inside a relational database and enable biological queries to be performed across the data. For example, which genes are differentially expressed in liver cancer or under which conditions is HOXA1 differentially expressed can be asked. Ontologies allow for much richer queries by additionally searching across synonyms entered in searches and by utilising the subsumption hierarchy within the ontology to return subclasses when a higher level superclass is used in the query (e.g. cancer returns experiments annotated with leukemia).

Understanding the role of genes within biological processes, such as diseases, is an important area of research [6]. Gene expression, however, is only one part of the larger biological picture. Once a list of genes of interest is produced, a

researcher will often wish to then investigate further by identifying, for example, which proteins, biological processes or signaling pathways correspond to the biological conditions of interest [7]. Typically this process will require the biologist to perform similar types of searches across multiple resources such as UniProt or Reactome before aggregating the resulting data and attempting to make sense of them.

In the post-genomic era, the heterogeneity of data models and formats has made interoperability and data integration difficult [8]. The emergence of bio-ontologies has improved this problem, although divergences across the ontology community also present their own challenges [9]. Despite these challenges, a wealth of ontologies now exist that are providing the necessary vocabulary for consistent annotation of a wide variety of biological data. These ontologies are readily available through resources such as the NCBO BioPortal [10], and the OBO foundry [11]. This availability has been key to the adoption of these ontologies to annotate and integrate resource and database like those hosted at the EBI.

Providing universal access to the often complex and heterogeneous data being generated in the life sciences continues to be a challenge in bioinformatics. Many databases now provide programmatic access to data through the use of web browsers and web service based APIs. These APIs are useful for exposing the data within a single resource, but integrating and querying across many resources is still problematic. Where ontological annotations are concerned, few resources exploit the full potential of the ontological structure to answer queries [12]. Moreover, the ability to ask semantically meaningful queries across resources such that one item of data is informed by data from another resource is either not possible, requires all relevant data to be aggregated into a single platform (e.g. Endeavour [13]) or requires very complex workflow solutions.

The Semantic Web promises solutions to many of the challenges of large scale data integration [14], and has subsequently received a lot of interest and adoption from the life sciences [15] [16]. At the core of the Semantic Web is the Resource Description Framework (RDF) [1], which provides a mechanism for publishing and describing data on the Web. RDF is a data model for describing graphs, where the semantics of relationships between nodes can be made explicit. RDF benefits from being built on existing Web technology. Uniform Resource Identifiers (URIs) provide a mechanism for identifying resources or data on the Web. HTTP provides a communication protocol for retrieving information about those resources, such as relationships to other resources. A common XML based exchange format for RDF exists along with a standardised query language called SPARQL [2]. Linking Web resources using RDF is at the core of the Linked Open Data movement which is a growing set of resources linked by RDF [17].

An increasing amount of life science datasets are becoming available as RDF, and publishing guidelines are emerging through community efforts such as the W3C Health Care and Life Science working group (HCLS) [18]. Projects such

[1] `http://www.w3.org/RDF`
[2] `http://www.w3.org/TR/rdf-sparql-query`

as Bio2RDF [19] provide access to a wide range of linked life science data and previous efforts such as the LODD dataset [20], and KUPKB [21] show how Semantic Web technology can enable novel biological discovery [22]. As a major bioinformatics service provider, the EBI is committed to adding value to existing resources. Providing production quality RDF that is synchronised with release cycles is a major step in that direction.

Two notable resources publishing RDF are the UniProt databases of protein sequence [23] and functions and the Reactome database of biological pathways [24] both developed in collaboration with the EBI. These two resources are already cross linked via UniProt accessions. Several other EBI resources are beginning to publish RDF, such as the ChEMBL database of bioactive drug-like small molecules [25], which is generating linked RDF as part of the OpenPhacts project to create a connected knowledge base of pharmacological data [26]. The ultimate goal is to produce an integrated set of production quality EBI resources published as RDF. In this paper we report on the development of the GXA database RDF resource and describe how we link to other EBI resources. Combining gene expression data with protein function information right through to the pathways where those proteins are active enables us to explore these datasets from multiple biological perspectives. We demonstrate some of the added value we gain from this RDF representation by describing some of the new queries we can now ask, show how the representation can improve error detection and illustrate some of the additional biological insights that can be gained by querying multiple integrated resources at once with an example relating to the study of diabetes.

2 Method

2.1 The GXA data

The GXA represents a subset of experiments from the ArrayExpress archive, that have data amenable to the GXA statistical analysis methodology, and meet a minimum criteria related to the numbers of replicates and ability to map the array design to a current genome build [2]. For each study the re-analysis of the raw expression data produces lists of differentially expressed genes relating to certain biological conditions. For each assay within an experiment, multiple biological conditions may be under investigation. Both the biological condition and sample descriptions are composed using a simple type/value notation. e.g. type = organism_part, value = liver. The GXA statistics are only generated for studies where the samples are annotated to more than one different biological condition e.g. liver and lung. The type/value pairs go through both a manual and semi-automated curation pipeline [3] in order to clean up and normalise the annotations (this is in addition to the curation that has already taken place within ArrayExpress). This data cleanup combined with the ontological markup

[3] `http://zooma.sourceforge.net`

adds value to the source data and enables tighter integration and more powerful queries.

We can conceptually separate the GXA model into three components that capture different aspects of the data. The first component is the notion of the study itself. Each study has associated meta-data such as the accession number, description, submitter, submission date and related publications. This information is derived from the original Investigation Description File (IDF) submitted to the AE archive, however, only a subset is retained for the GXA. We initially focus on capturing only this subset in RDF. A complete conversion of the AE data, and microarray data in general, into RDF is beyond the scope of this work and the subject of other ongoing work [27].

Each experiment has a set of associated assays. Each assay has an accession number, and links to a set of experimental factors and may also link to related sample information. Each factor and sample has a type and value annotations, along with any optional ontological annotations. Figure 1 shows the basic model for capturing assay and sample information. This information originates from the Sample and Data Relation File (SDRF) originally associated with the AE submission. This information is subjected to manual and semi-automated curation to normalise property types, and apply ontological annotation to property values using the Experimental Factor Ontology (EFO).

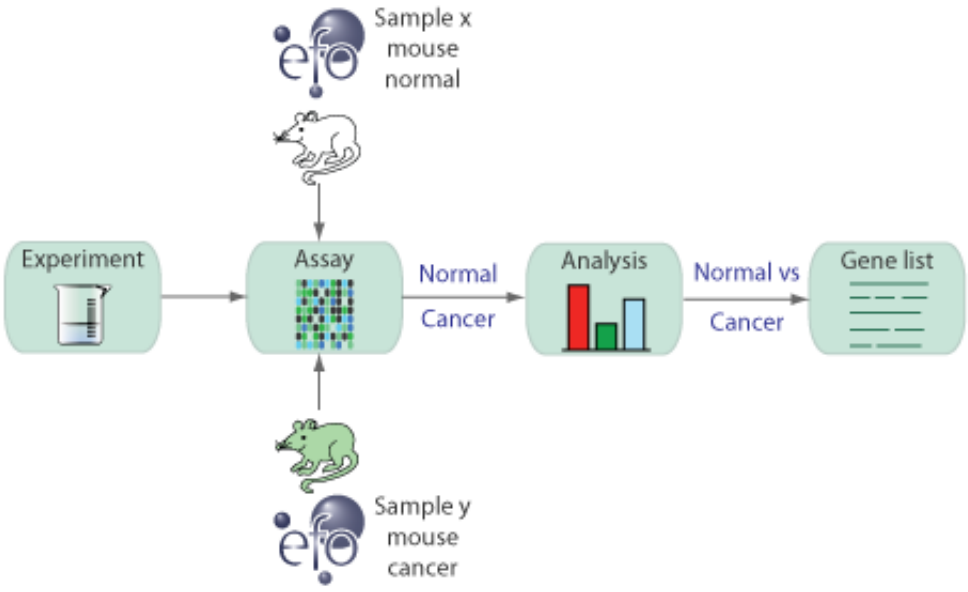

Fig. 1. Abstract model for the data in GXA. Experiments are linked to samples, that are used in some analysis to compute gene lists.

Finally, each experiment can have multiple data analyses associated with it, each of which produces a set of gene lists. The analysis should record the provenance of when the analysis was done, which software was used and the set of input files used. Some studies have more than one type of array design, so separate gene lists are computed for each platform within the study. A gene list is generated for each value of a biological condition for a given type (e.g. values cisplatin and untreated for type compound). For each list one of the biological condition values is acting as the independent variable. All remaining assays for that condition type play the role of the control.

For example, after analysis, experiment E-GEOD-24868 [4] has two gene lists. The first is differentially expressed genes where the independent variable is cell type / PC-3/S, and a second gene list where the independent variable is cell type / PC-3/Mc. The gene lists includes a reference to the probe sets, or design elements, for the given platform, along with a measure of differential gene expressions and an associated confidence measure. In order to map the probe ID to a given gene, the GXA generates mappings to Ensembl genes at every release.

2.2 Ontologies

The GXA already uses EFO as the primary annotation ontology for biological conditions and bio-sample information [5]. EFO is an application ontology that imports terms from a set of reference ontologies based on the criteria below, and defines terms locally where a suitable term cannot be found in an external ontology that meets these criteria:

- Coverage - that the ontologies include classes that describe concepts in the experimental data;
- That are available in OWL or OBO format;
- That are maintained or are actively developed;
- Have some evidence of usage across the community.

Figure 1 provides an outline for the RDF graph needed for modeling GXA data. Using the ontology selection criteria above, we used the NCBO bioportal to survey a wide range of ontologies for typing resources in the GXA model. We found considerable overlap for many of the terms, especially for more general concepts such as experiment, assay, sample, gene list and gene expression. The Ontology of Biomedical Investigations (OBI) [28] provides good term coverage for microarray experiments due to the inclusion of many terms from the original MGED ontology [29]. Despite the good term coverage, the OBI model is complex when it came to asserting relations between concepts. OBI provides a framework for creating very accurate semantic descriptions of any type of biomedical investigation. As such it has a very large, overarching framework to satisfy the very wide range of queries that may be possible about an investigation for any domain. This often involves making many distinctions and assertions that add

[4] `http://identifiers.org/gxa.expt/E-GEOD-24868`

complexity to the model but add little extra value in the context of the narrower queries needed in this work, such as distinguishing between a written plan specification and the duplicated execution of this plan in a process as separate entities. These descriptions add considerable complexity to the underlying RDF model that makes querying and exploring the data harder.

Instead, a lighter weight schema, that makes fewer ontological commitments, but has sufficient semantics to satisfy our collected competency questions was required. As such, some of the more general purpose vocabularies were also sourced to provide predicates that capture relationships between data. In addition to OBI, we looked at the OBO relation ontology [30], the Information Artefact Ontology (IAO) [5], Semantic-science Integration Ontology (SIO) [6] and PROV-O [7]. There was significant overlap between several of these ontologies although only SIO provided all of the classes and predicates required for capturing the model required for the GXA RDF. In order to maximise external integration, we created equivalence statements between classes and predicates which overlapped and incorporated both into the RDF store. The schema is currently available at `http://wwwdev.ebi.ac.uk/fgpt/efosemweb.html`.

It is worth noting that there is ongoing work as part of the W3C HCLS group to provide a best practice note on describing gene expression data in RDF but that this is presently incomplete. Our initial focus for this work is to ask novel queries of the existing GXA data rather than to produce a model by which all microarray or sequencing data should be described.

2.3 Identifiers

A URI scheme was devised for the GXA to ensure that the data conforms to Linked Open Data guidelines [8] Best practice for linked data state that you should:

- use URIs to name things
- use HTTP URIs so that people can look up those names
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs. so that they can discover more things.

We address issues 1, 2 and 3 by generating URIs using the identifiers.org [31] registry, hosted at the EBI. Identifiers.org provides resolvable persistent URIs which can be set to redirect HTTP request to the relevant resource. This mechanism will ensure that the URIs used in the GXA RDF will always resolve to some resource and removes any dependency between GXA RDF URIs and the GXA web server.

[5] `http://code.google.com/p/information-artifact-ontology`
[6] `http://code.google.com/p/semanticscience`
[7] `http://www.w3.org/TR/prov-o`
[8] `http://www.w3.org/DesignIssues/LinkedData.html`

All external terminologies and ontologies referenced by the GXA RDF model conform to the same set of linked data principles, which ensures we satisfy criteria 4. In addition to linking out to other resources relating to aspects of the study and study design, we also wanted to link the gene lists to relevant resources. GXA currently maps probe set ids to gene accessions from the Ensembl genome database. Ensembl are yet to publish their data in RDF, however, they have registered a URI scheme with identifiers.org. We can now link out from GXA genes to Ensembl URIs using the Identifers.org schema, thus future proofing our links.

Given the central importance of a gene to so much of the study of biology, it naturally provides us with a vast amount of linking potential from the transcriptomics data described in the GXA. Another important dataset is the UniProt database, which provides a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. UniProt already publish their data as RDF and have a persistent URL strategy based on PURLs [9]. As part of the GXA RDF build, we generate simple RDF datasets that provides links from Ensembl genes to their associated UniProt Proteins. These mappings are derived from the monthly UniProt release available from the UniProt FTP (`ftp://ftp.uniprot.org/pub/databases/uniprot`). Linking to UniProt enables us to link the GXA data out into other RDF datasets such as the Gene Ontology Annotations (GOA) for protein function, Reactome for pathway data, and ChEMBL for chemical compounds.

2.4 Data processing

The pipeline for building the GXA RDF is written in Java and uses the OpenRDF Sesame framework [32], the OWL-API [33] and GXA API [10]. Upon each release of the GXA, the API is used to read data out of the GXA database, each experiment is processed and converted into an individual OWL/RDF file. The file is classified using an OWL reasoner to precompute any additional entailments. Once all the files have been generated a separate pipeline is used to load each file into a triple store. We are currently using the OWLIM-SE [34] version 5 triple store within the OpenRDF Sesame framework. The full pipeline is set to run on the EBIs computing cluster and currently takes around 18 hours to complete.

2.5 User group competency questions

User engagement is a key component of this work. In order to understand how we could usefully extend the Gene Expression Atlas, we engaged in a series of user interviews to elicit requirements. These requirements were shaped into the form of competency questions which we would build the representation towards answering and which would then form a set of evaluation criteria which we could

[9] `http://purl.oclc.org`
[10] `https://github.com/gxa/gxa`

test the resource with. These users were from a diverse set of backgrounds and held a varying array of expertise which included mouse phenotype biologists, bacteria experimentalists, data analysts and clinicians.

Examples of questions included the following:

- Which genes are differentially expressed in adult mice bred in oxygen rich vs oxygen poor environments?
- Which genes with antigen binding function are differentially expressed, where and with regard which diseases and which pathways?
- Which genes involved in Basigin interactions are differentially expressed in Malaria?
- Which genes are differentially expressed in a knockout mouse strain versus the wildtype?
- Which genes are differentially expressed in diseases which occur in the urinary tract?

3 Results

Using the described model, an RDF store of data from version 12.7 of the Gene Expression Atlas at EBI was produced. This included a total of 3,317 experiments which is the majority of the current Gene Expression Atlas and covered 328,645 genes across 83,233 assays. 28,702 gene lists were described as a result of the transformation into the RDF model. The total number of triples loaded is 773,025,372. The generated RDF files will soon be available to download, and are already available via a public SPARQL endpoint [11].

We have included mappings between probeset IDs, Ensembl identifiers and UniProt identifiers. We loaded version 40 of the Reactome RDF database so that we can integrate queries across UniProt accessions. Identifiers.org provided us with URIs to reference Ensembl genes, which provide us with a large amount of integration potential. Finally we can integrate gene products to UniProt accessions held in the UniProt SPARQL endpoint [12]. Examples are given on the GXA RDF website [13] that demonstrate for the first time federated queries across the GXA and UniProt resources. For example, we can now query for genes differentially expressed and filter according to the gene product GO annotations from UniProt. The use of identifers.org to mint URIs for GXA resource, in addition to our efforts to link out to other RDF resources puts the GXA firmly in the

[11] `http://wwwdev.ebi.ac.uk/fgpt/gxa-sparql`
[12] `http://beta.sparql.uniprot.org`
[13] `http://wwwdev.ebi.ac.uk/fgpt/gxa-sparql`

growing Linked Open Data cloud. More importantly the rich biological information held in GXA serves as a valuable addition to the set of life science datasets already published as RDF.

3.1 Integrating across resources

One of the objectives of this work was to enable querying across a wider range of biological resources than is currently possible with the GXA, with particular reference to queries required by our target user group. Our RDF model for describing GXA results enables queries such as "*find differentially expressed genes in liver cancer versus normal tissue*" to be framed, which was not previously possible. In addition, with the expressiveness granted by including the EFO within a framework native to RDF, queries which utilise some of the axioms can now be expressed. For example, find differentially expressed genes which have disease location in genitourinary system which calls upon the axioms within EFO using the object property has_disease_location to link disease to organism parts. A further type of query is when we wish to be explicit about which results we do not want to return. For example, querying for experiments in which the independent variable is not disease rather than asking for independent variables that are annotated with normal. It is noteworthy that these are two different queries, even though they would seem identical. Normal is often used to describe some background normality within the experiment and is therefore contextual; this does not necessarily imply it is free from disease (or never was). We can see examples of this in multiple experiments, such as E-TABM-1171, in which the experiment is described as differential expression in cancer samples compared to normal samples. Here, the normal sample is annotated as normal and also benign; the annotation should really be interpreted as not cancer. With the GXA RDF store we can ask this query making use of the NOT EXISTS filter in SPARQL, by simply asking for independent variables which are not of the type disease.

3.2 Answering biologically relevant questions

Our engagement with user groups, as described in the previous section, produced a list of queries that were desirable of the new GXA RDF store. In order to evaluate these competency questions, we first designed SPARQL queries intended to answer these questions and secondly, performed an evaluation of the results using the literature. Examples of these queries can be seen on the].

One category of query often requested by users is the ability to view differentially expressed genes in the context of pathways. It is known that most phenotypes observed are the result of combination of genes. When studying a particular phenomenon for gene expression, techniques such as enrichment analysis exploit known information about individual genes to see if if groups of related genes are being differentially expressed. These groups are typically categories taken from the Gene Ontology, but similar analysis can be performed using alternate gene annotation, such as associated pathways. We can begin to explore this kind of

query using the GXA RDF by exploiting the new links between gene expression and the Reactome pathways data.

One query of interest to our biologist user group concerned pathways in which genes were differentially expressed in muscle tissues where the sample is taken from an experiment studying diabetes vs normal. We can construct a SPARQL query to get all differentially expressed genes in muscle where the experimental factor is diabetes. This initial query returns 287 genes. We can now refine this query to only include genes with known human pathway associations from Reactome, which reduces the set of genes to just 31. From that list of genes we see experiment E-GEOD-1659 (which is investigating gene expression in mice muscle before and after exercise) has found differential expression in one of those genes - the Receptor (calcitonin) activity modifying protein 1 (Ramp 1 [14]) being down regulated in mice. The query also tells us that Ramp1 has an association to the Calcitonin-like ligand receptors pathway [15]. A short literature survey confirms the association of Ramp genes to the Calcitonin pathway. It also suggests evidence for associations between this pathway to energy homeostasis, obesity and diabetes [35] [36]. What these results demonstrate is that, even though the experiments are studying different conditions (diabetes in one, exercise in another), they report similar genes which are involved in similar pathways and that literature can confirm that this pathway may have roles to play in all of these conditions. The data contains such information but it is only made apparent once everything is integrated together. Where literature does not confirm such associations, these can be considered potential hypothesis for further investigation.

Scanning the list of 31 candidate genes and pathways from the previous query requires a trained eye from a scientist working in the domain in order to assess whether they have any relevance. They key point is that the integration is working to bring relevant information together and provide alternate facets over the data. In this case we facet the results by pathways, but we could easily link out to UniProt to retrieve GO annotations for protein function and facet along those. This kind of data exploration can enable potentially new hypotheses generation or can help to validation known or expected information, show contradictory information or simply open up new avenues for exploration.

3.3 Annotation coverage and error detection

One of the benefits of the RDF model is the ability to ask questions about individuals relating to their type. An example of where this is particularly useful is when looking for possible annotation inconsistency or error. For the factor type 'disease_state, for example, we would expect to only see annotations which are types of disease. To test this, we performed a SPARQL query for factor value individuals which are not types of disease where the factor type was 'disease_state. The results of this query reveal two findings. Firstly, that there are annotations

[14] `http://identifiers.org/ensembl/ENSMUSG00000034353`
[15] `http://www.ncbi.nlm.nih.gov/biosystems/106379`

of the type 'normal and 'normospermic' (normal sperm levels) annotated under 'disease_state' within the data; this may or may not be an error in annotation, since, as previously discussed normal is contextual - but it is arguable that normal should not be considered a disease state. Secondly, it highlights annotations which are not types of disease but should be, that is gaps in coverage in the ontology. For example, 'rectovaginal endometriosis' and 'Juvenile Idiopathic Scoliosis' were two such annotations both of which represent missing coverage in the ontology.

4 Conclusion

The primary goal of this work was to enable new biologically interesting queries to be asked of the data in the Gene Expression Atlas by integration with external resources including protein databases and pathways. This work demonstrates that a lot can be gained from relatively simple integration between resources and, moreover, that the key biological entities of genes and proteins can be seen as anchors which connect a lot of biomedical data. The addition of ontologies such as EFO and GO allow richer, more expressive queries to be asked of the data - made possible because the data are curated with such ontologies, thereby demonstrating the power of well-annotated data.

The choice of model here is relatively minimal and reuses existing concepts where possible. Our initial approach has been to import a lot of the data into the RDF store but the longer term aim is to use federated querying to integrate these data. New SPARQL endpoints at EBI for Genome Wide Association Study (GWAS), ChEMBL and BioModels, along with the offering from the NCBO will all provide more integration points. The GXA RDF is a starting point towards a more integrated RDF offering from resources at EBI. Our approach to this is agile; we develop in small iterations, document and refine with users involved at each stage. In this respect, engineering RDF is no different from good software engineering practices.

Future work will include extending the model to capture more general information about a microarray experiment. The end goal will be to provide RDF data for all experiments in the ArrayExpress archive. This will require more engagement with the community through efforts such as the HCLS, to generate agreement on the model and terminology. In particular, through our integration with the SIO we would like to explore how we can expose our service to semantic service platform such as SADI [37].

Whilst SPARQL provides us with a low level query language for exploring and mining the data, this form of advanced API will only be directly useful to a small subset of our users. The next step is to develop new user interface components that allow the user to exploit these rich data connections, whilst at the same time shielding the user from the underlying technology. Previous work [38] has shown that without developing user facing tools for biologist, much of the added value in the RDF is not available to most users. We are currently developing tools to do gene set enrichment analysis based on the GXA annotations and

an improved user interface to allow better interaction with our biologist target audience.

5 Acknowledgements

We acknowledge funds from EMBL (JM, HP) and The National Center for Biomedical Ontology, one of the National Centers for Biomedical Computing supported by the NHGRI, the NHLBI, and the NIH Common Fund under grant U54-HG004028 (SJ). We are also grateful for discussion and comments from Tony Burdett, Maryam Soleimani, Johan Rung at the EBI, Robert Petryszak, Eleanor Williams and Maria Keays and the Gene Expression Atlas production team, and Michel Dumontier and members of the HCLS interest group.

References

1. Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Rustici, G., Williams, E., Parkinson, H., Brazma, A.: Gene expression atlas at the european bioinformatics institute. Nucleic Acids Research **38**(suppl 1) (2010) D690–D698
2. Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N., Kurnosov, P., Malone, J., Melnichuk, O., Petryszak, R., Pultsin, N., Rustici, G., Tikhonov, A., Travillian, R.S., Williams, E., Zorin, A., Parkinson, H., Brazma, A.: Gene expression atlas updatea value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Research **40**(D1) (2012) D1077–D1081
3. Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N., Brazma, A.: Arrayexpress updatean archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Research **39**(suppl 1) (2011) D1002–D1004
4. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Khri, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernndez-Suarez, X.M., Harrow, J., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., Searle, S.M.J.: Ensembl 2012. Nucleic Acids Research **40**(D1) (2012) D84–D90
5. Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H.: Modeling sample variables with an experimental factor ontology. Bioinformatics **26**(8) (2010) 1112–1118
6. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J.,

Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**(6871) (January 2002) 530–536
7. Nam, D., Kim, S.Y.: Gene-set approach for expression pattern analysis. Briefings in Bioinformatics **9**(3) (2008) 189–197
8. Antezana, E., Kuiper, M., Mironov, V.: Biological knowledge management: the emerging role of the semantic web technologies. Briefings in Bioinformatics **10**(4) (2009) 392–407
9. Lord, P., Stevens, R.: Adding a little reality to building ontologies for biology. PLoS ONE **5**(9) (09 2010) e12258
10. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Research **39**(suppl 2) (2011) W541–W545
11. Smith, Barry, Ashburner, Michael, Rosse, Cornelius, Bard, Jonathan, Bug, William, Ceusters, Werner, Goldberg, Louis J., Eilbeck, Karen, Ireland, Amelia, Mungall, Christopher J., Leontis, Neocles, Rocca-Serra, Philippe, Ruttenberg, Alan, Sansone, Susanna-Assunta, Scheuermann, Richard H., Shah, Nigam, Whetzel, Patricia L., Lewis, Suzanna: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology **25**(11) (November 2007) 1251–1255
12. Jupp, S., Stevens, R., Hoehndorf, R.: Logical gene ontology annotations (goal): exploring gene ontology annotations with owl. Journal of Biomedical Semantics **3**(Suppl 1) (2012) S3
13. Tranchevent, L.C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., Moreau, Y.: Endeavour update: a web resource for gene prioritization in multiple species. Nucleic Acids Research **36**(suppl 2) (2008) W377–W384
14. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5) (2001) 34–43
15. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the semantic web. BMC Bioinformatics **8**(Suppl 3) (2007) S2
16. Cheung, K.H., Yip, K.Y., Smith, A., deKnikker, R., Masiar, A., Gerstein, M.: Yeasthub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics **21**(suppl 1) (2005) i85–i96
17. Nolin, M.A., Dumontier, M., Belleau, F., Corbeil, J.: Building an hiv data mashup using bio2rdf. Briefings in Bioinformatics **13**(1) (2012) 98–106
18. Marshall, M.S., Boyce, R., Deus, H.F., Zhao, J., Willighagen, E.L., Samwald, M., Pichler, E., Hajagos, J., Prud'hommeaux, E., Stephens, S.: Emerging practices for mapping and linking life sciences data using rdf a case series. Web Semantics: Science, Services and Agents on the World Wide Web **14**(0) (2012) 2 – 13 Special Issue on Dealing with the Messiness of the Web of Data.
19. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics **41**(5) (2008) 706 – 716 Semantic Mashup of Biomedical Data.

20. Luciano, J., Andersson, B., Batchelor, C., Bodenreider, O., Clark, T., Denney, C., Domarew, C., Gambet, T., Harland, L., Jentzsch, A., Kashyap, V., Kos, P., Kozlovsky, J., Lebo, T., Marshall, S., McCusker, J., McGuinness, D., Ogbuji, C., Pichler, E., Powers, R., Prud'hommeaux, E., Samwald, M., Schriml, L., Tonellato, P., Whetzel, P., Zhao, J., Stephens, S., Dumontier, M.: The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. Journal of Biomedical Semantics **2**(Suppl 2) (2011) S1
21. Jupp, S., Klein, J., Schanstra, J., Stevens, R.: Developing a kidney and urinary pathway knowledge base. Journal of Biomedical Semantics **2**(Suppl 2) (2011) S7
22. Klein, J., Jupp, S., Moulos, P., Fernandez, M., Buffin-Meyer, B., Casemayou, A., Chaaya, R., Charonis, A., Bascands, J.L., Stevens, R., Schanstra, J.P.: The kupkb: a novel web application to access multiomics data on kidney disease. The FASEB Journal **26**(5) (2012) 2145–2153
23. Consortium, T.U.: Ongoing and future developments at the universal protein resource. Nucleic Acids Research **39**(suppl 1) (2011) D214–D219
24. Croft, D., OKelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., DEustachio, P., Stein, L.: Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Research **39**(suppl 1) (2011) D691–D697
25. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P.: Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Research **40**(D1) (2012) D1100–D1107
26. Harland, L.: Open phacts: A semantic knowledge infrastructure for public and commercial drug discovery research. In Teije, A., Vlker, J., Handschuh, S., Stuckenschmidt, H., dAcquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N., eds.: Knowledge Engineering and Knowledge Management. Volume 7603 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 1–7
27. Deus, H.F., Prudhommeaux, E., Miller, M., Zhao, J., Malone, J., Adamusiak, T., McCusker, J., Das, S., Serra, P.R., Fox, R., Marshall, M.S.: Translating standards into practice one semantic web api for gene expression. Journal of Biomedical Informatics **45**(4) (2012) 782 – 794 ¡ce:title¿Translating Standards into Practice: Experiences and Lessons Learned in Biomedicine and Health Care¡/ce:title¿.
28. Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A.A., Soldatova, L.N., Stoeckert, C.J., Turner, J.A., Zheng, J., OBI consortium: Modeling biomedical experimental processes with OBI. Journal of biomedical semantics **1 Suppl 1**(Suppl 1) (2010) S7+
29. Whetzel, P.L., Parkinson, H., Causton, H.C., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., Sansone, S.A., Taylor, C., White, J., Stoeckert, C.J.: The mged ontology: a resource for semantics-based description of microarray experiments. Bioinformatics **22**(7) (2006) 866–873
30. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., Rosse, C.: Relations in biomedical ontologies. Genome Biology **6**(5) (2005) R46
31. Juty, N., Le Novre, N., Laibe, C.: Identifiers.org and miriam registry: community resources to provide persistent identification. Nucleic Acids Research **40**(D1) (2012) D580–D586

32. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A generic architecture for storing and querying rdf and rdf schema. In Horrocks, I., Hendler, J., eds.: The Semantic Web - ISWC 2002. Volume 2342 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2002) 54–68
33. Horridge, M., Bechhofer, S.: The owl api: A java api for owl ontologies. Semantic Web **2**(1) (2011) 11–21
34. Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., Velkov, R.: Owlim: A family of scalable semantic repositories. Semant. web **2**(1) (January 2011) 33–42
35. Zhang, Z., Liu, X., Morgan, D.A., Kuburas, A., Thedens, D.R., Russo, A.F., Rahmouni, K.: Neuronal receptor activitymodifying protein 1 promotes energy expenditure in mice. Diabetes **60**(4) (2011) 1063–1071
36. Bailey, R., Walker, C., Ferner, A., Loomes, K., Prijic, G., Halim, A., Whiting, L., Phillips, A., Hay, D.: Pharmacological characterization of rat amylin receptors: implications for the identification of amylin receptor subtypes. British Journal of Pharmacology **166**(1) (2012) 151–167
37. Wilkinson, M.D., Vandervalk, B.P., McCarthy, E.L.: Sadi semantic web services – 'cause you can't always get what you want!. In Kirchberg, M., Hung, P.C.K., Carminati, B., Chi, C.H., Kanagasabai, R., Valle, E.D., Lan, K.C., Chen, L.J., eds.: APSCC, IEEE (2009) 13–18
38. Jupp, S., Klein, J., Moulos, P., Schanstra, J., Stevens, R.: Ontologies come of age with the ikup browser. In: Proceedings of the Workshop Ontologies Come of Age in the Semantic Web, International Semantic Web Conference. (2011)