Simplifying Access to Large-Scale Health Care and Life Sciences Datasets

Holger Stenzhorn[1,2], Kavitha Srinivas[3], Matthias Samwald[2,4], and Alan Ruttenberg[5]

[1] Department of Medical Informatics, University Medical Center Freiburg, Germany
[2] Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Ireland
[3] IBM T.J. Watson Research Center, Yorktown Heights, New York, USA
[4] Section on Medical Expert and Knowledge-Based Systems, Medical University of Vienna, Austria
[5] Science Commons, Cambridge, Massachusetts, USA

1 Introduction

Within the health care and life sciences (HCLS) domain, a plethora of phenomena exists that range across the whole "vertical scale" of biomedicine. To accurately research and describe those phenomena, a tremendous amount of highly heterogeneous data have been produced and collected with various research methodologies encompassing the genetic, molecular, tissue, and organ level. An initial step to provide researchers with access to this data has been through creating integrated views on existing and open biomedical datasets published on the Web. In order to make the next step, we need to now create easy-to-use yet powerful applications that enable researchers to efficiently query, integrate and analyze those datasets.

One effort in that direction is currently carried out by the World Wide Web Consortium's Semantic Web Health Care and Life Sciences Interest Group (HCLSIG)[1]. It is intended as a bridge between the Semantic Web community's technology and expertise and the information challenges and experiences in the HCLS communities [8]. It brings together scientists, medical researchers, science writers, and informaticians working on new approaches to support biomedical research. Participants come from both academia, government, non-profit organizations as well as health care, pharmaceuticals, and industry vendors.

In the following we show some results of this effort by describing two demonstrations of our approach on preparing and applying biomedical information on the Semantic Web. (All demonstration materials can be freely downloaded[2].)

2 Data Modeling, Storage and Provision

As starting point for our activities, we have re-modeled several biomedical datasets in OWL in order to take advantage of that language's well-defined semantics.

[1] `http://www.w3.org/2001/sw/hcls`
[2] `http://esw.w3.org/topic/HCLS/Banff2007Demo`

S. Bechhofer et al.(Eds.): ESWC 2008, LNCS 5021, pp. 864–868, 2008.

Those datasets include[3] PubMed[4], the Gene Ontology Annotations (GOA)[5], Entrez Gene[6], the Medical Subject Headings (MeSH)[7], the Foundational Model of Anatomy (FMA)[8], and the Allen Brain Atlas (ABA)[9].

2.1 Modeling Principles

Great care has been taken in modeling the dataset, e.g., to clearly distinguish between database records and real world statements such as about proteins in cells. We hase used (and extended) the OBO Foundry design principles [5] to create interoperability between the information sources and the OBO ontologies[10].

As a proof for the viability of this approach, we have successfully aligned a principled representation of GOA with two new representations of neuroscience databases (NeuronDB[11] and the Brain Architecture Management System (BAMS)[12]). Additionally, to create a specific anatomical view over the resources we have created mappings from MeSH to the FMA using UMLS[13].

2.2 Storage and Provision

We have created two demonstrations based on the OWL dataset representations:

In the first one, we have used the Openlink Virtuoso[14] RDF triple store to save more than 300 million triples (made publicly accessible through a SPARQL endpoint[15]). Although it is possible to store OWL in triple stores, Virtuoso does not support native OWL inference. In order to support more expressive queries, we have provided limited inference support in this implementation by using a combination of Virtuoso's native transitive closure support, simple rules based on their implementation of SPARQL-Update (SPARUL)[16] and loading partonomy relationships[17] pre-computed by the Pellet OWL Reasoner [6].

In the second one, we have used the SHER OWL management and inference system from IBM Research [2]. In this system also about 300 million triples have been stored, providing OWL inferencing over both GO[18] and FMA [4].

[3] The full dataset list is available at `http://sw.neurocommons.org/2007/kb-sources`
[4] `http://www.ncbi.nlm.nih.gov/pubmed`
[5] `http://www.geneontology.org/GO.annotation.shtml`
[6] `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`
[7] `http://www.nlm.nih.gov/mesh`
[8] `http://sig.biostr.washington.edu/projects/fm`
[9] `http://www.brainatlas.org`
[10] `http://www.obofoundry.org`
[11] `http://senselab.med.yale.edu/senselab/NeuronDB`
[12] `http://brancusi.usc.edu/bkms`
[13] `http://umlsinfo.nlm.nih.gov`
[14] `http://www.openlinksw.com/virtuoso`
[15] `http://hcls.deri.ie/demo`, `http://sparql.neurocommons.org:8890/nsparql`
[16] `http://jena.hpl.hp.com/~afs/SPARQL-Update.html`
[17] `http://esw.w3.org/topic/HCLS/PartOfInference`
[18] `http://www.geneontology.org`

2.3 Identifiers

We have also assured to employ a URI scheme to uniquely name resources and biological entities based on the `purl.org` resolver: Stable URIs have been given both to existing resources where providers do not currently have any stable identification scheme as well as for newly defined classes and instances. To avoid resolver redirection overhead for large query numbers, Semantic Web agents can query the resolver once to retrieve rewrite rules and implement those in their application to then access the actual resource directly.

3 Query Capabilities

Our two implementations highlight two different approaches to inference and reasoning. But both of them aim at retrieving precise answers to narrow queries.

Fig. 1 shows an example of a SPARQL query against our triple store querying for *genes associated with CA1 Pyramidal Neurons* (as defined by MeSH) and *signal transduction processes* (as defined by GO), returning 40 pairings of gene and process, compared to about 175,000 returned by the query against Google for *genes involved in pyramidal neuron signal transduction*.

```
prefix go: <http://purl.org/obo/owl/GO#>
prefix mesh: <http://purl.org/commons/record/mesh/>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix sc: <http://purl.org/science/owl/sciencecommons/>
prefix ro: <http://www.obofoundry.org/ro/ro.owl#>

SELECT DISTINCT ?gene ?process WHERE {
  graph <http://purl.org/commons/hcls/pubmesh>
  { ?pubmedrecord ?p mesh:D017966.
    ?article sc:identified_by_pmid ?pubmedrecord.
    ?generecord sc:describes_gene_or_gene_product_mentioned_by ?article. }
  graph <http://purl.org/commons/hcls/goa>
  { ?protein rdfs:subClassOf ?res.
    ?res owl:onProperty ro:has_function.
    ?res owl:someValuesFrom ?res2.
    ?res2 owl:onProperty ro:realized_as.
    ?res2 owl:someValuesFrom ?process.
    graph <http://purl.org/commons/hcls/20070416/classrelations>
      {{ ?process <http://purl.org/obo/owl/obo#part_of> go:GO_0007166. }
         union { ?process rdfs:subClassOf go:GO_0007166. }}
    ?protein rdfs:subClassOf ?parent.
    ?parent owl:equivalentClass ?res3.
    ?res3 owl:hasValue ?generecord. }}
```

Fig. 1. Complex SPARQL query to retrieve all genes which are associated with both *CA1 Pyramidal Neurons* and the *signal transduction processes*

The query works by linking MeSH associated with Pubmed records to genes via Entrez Gene, narrowing the genes by GO associations narrowed to signal transduction processes or parts of those processes.

On the other hand, SHER provides reasoning capabilities over FMA and GO. Reasoning on FMA is well known to be problematic for current reasoners due to the fact that FMA represents a deep mereological hierarchy in which both *part-of* as well as its inverse *has-part* relations are employed [1]. This occurs partly to work around modeling constraints found in OWL. Even though mereological hierarchies are better modeled as description graphs [3] this would require a complete re-modeling of FMA, and hence, as a workaround, we reasoned only over the *part-of* relations in FMA with the SHER OWL reasoner. Supported sample queries are of the following form (marked concepts are taken from FMA and GO respectively and require inferencing): *Find the genes known to be involved in Alzheimer's disease, in the* ***hippocampal*** *region that have a role in* ***dendrite development****.* This expands the search not only to the hippocampus but also to its sub-parts, such as the CA1 region as well as to processes that are part of dendrite development such as dendrite morphogenesis.

4 User Interface

In order to demonstrate the capabilities of the developed system we have created two different browser interfaces using freely available tools:

In the first one, we have combined query results with data made available by the Alan Brain Institute and presented the combined data using Exhibit [7]. It shows images of mouse brain slices stained for expressed genes with each gene's details, and visualizing its transcript regions and genomic context (cf. Fig. 2).

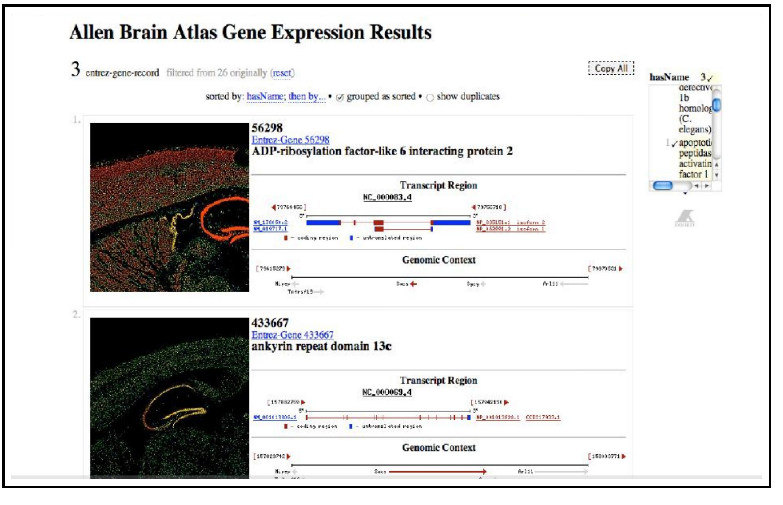

Fig. 2. Screenshot of a gene query result in Exhibit, showing expressions with images from the Allen Brain Atlas combined with transcripts from Entrez Gene

In the second interface, we have provided an intuitive keyword interface to search medical literature with keywords being internally converted into a logical query. The example sentence would be translated into (y is the selected variable) `aboutGene(x,y)` ⊓ `hasFunction(x,m)` ⊓ `rdf:type(m,GO:dendrite_development)` ⊓ `evidence(x,z)` ⊓ `source(z,p)` ⊓ `hasPubMedID(p,q)` ⊓ `hasAsMesh(q, Alzheimer's_disease)` ⊓ `hasAsMesh(q,r)`⊓ `rdf:type(r, FMA:hippocampus)`.

As a next step, we envisage to develop a further interface for simplifying the creation and maintainance of complex SPARQL queries for the first system.

Acknowledgments

This work is the product of many participants in the HCLSIG and includes (besides the authors) John Barkley, Olivier Bodenreider, Bill Bug, Huajun Chen, Paolo Ciccarese, Kei Cheung, Tim Clark, Don Doherty, Julian Dolby, Kerstin Forsberg, Achille Fokoue, Ray Hookaway, Aditya Kalyanpur, Vipul Kashyap, June Kinoshita, Joanne Luciano, Li Ma, Scott Marshall, Chris Mungall, Eric Neumann, Chintan Patel, Eric Prud'hommeaux, Jonathan Rees, Edith Schonberg, Mike Travers, Gwen Wong and Elizabeth Wu. Susie Stephens coordinated the BioRDF subgroup of the HCLSIG in which this work was developed.

References

1. Dameron, O., Rubin, D., Musen, M.: Converting the Foundational Model of Anatomy into OWL. In: Proc. AMIA Symp 2005, Washington, DC (2005)
2. Dolby, J., Fokoue, A., Kalyanpur, A., Kershenbaum, A., Ma, L., Schonberg, E., Srinivas, K.: Scalable Semantic Retrieval Through Summarization and Refinement. In: Proc. AAAI 2007, Vancouver, Canada (2007)
3. Motik, B., Cuenca Grau, B., Sattler, U.: Structured Objects in OWL: Representation and Reasoning. Technical Report, University of Oxford, UK (2007)
4. Fokoue, A., Kershenbaum, A., Ma, L., Schonberg, E., Srinivas, K.: The Summary Abox: Cutting Ontologies Down to Size. In: Proc. 5th International Semantic Web Conference, Athens, GA, USA (2006)
5. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Scheuermann, R., Shah, N., Whetzel, P., Lewis, S.: The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nature Biotechnology 25, 1251–1255 (2007)
6. Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. UMIACS Technical Report, 2005-68 (2005)
7. Huynh, D., Karger, D., Miller, R.: Exhibit: Lightweight Structured Data Publishing. In: Proc. of the World Wide Web 2007 Conference, Banff, Canada (2007)
8. 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., Ogbuji, C., Rees, J., Stephens, S., Wong, G., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.: Advancing translational research with the Semantic Web. BMC Bioinformatics 8 (2007)