

Using semantic web tools to integrate experimental measurement data on our own terms

M. Scott Marshall¹, Lennart Post^{1,2}, Marco Roos¹, Timo M. Breit¹

1. Integrative Bioinformatics Unit, 2. Nuclear Organisation Group
Institute for Informatics, Swammerdam Institute for Life Sciences
Faculty of Science
University of Amsterdam
marshall@science.uva.nl

Abstract. The -omics data revolution, galvanized by the development of the web, has resulted in large numbers of valuable public databases and repositories. Scientists wishing to employ this data for their research are faced with the question of how to approach data integration. Ad hoc solutions can result in diminished generality, interoperability, and reusability, as well as loss of data provenance. One of the promising notions that the Semantic Web brings to the life sciences is that experimental data can be described with relevant life science terms and concepts. Subsequent integration and analysis can then take advantage of those terms, exposing logic that might otherwise only be available from the interpretation of program code. In the context of a biological use case, we examine a general semantic web approach to integrating experimental measurement data with Semantic Web tools such as Protégé and Sesame. The approach to data integration that we define is based on the linking of data with OWL classes. The general pattern that we apply consists of 1) building application-specific ontologies for “myModel” 2) identifying the concepts involved in the biological hypothesis, 3) finding data instances of the concepts, 4) finding a common domain to be used for integration, and 5) integrating the data. Our experience with current tools indicates a few semantic web bottlenecks such as a general lack of ‘semantic disclosure’ from public data resources and the need for better ‘interval join’ performance from RDF query engines.

Introduction

The -omics data revolution, galvanized by the development of the web, has produced large numbers of valuable public databases and repositories. These databases enable many types of research by providing free web access to essential up-to-date -omics information and even raw data. However, the same revolution has also led to an explosion of proprietary formats and interfaces. Researchers who want to integrate data from several sources must find a way to extract information from a variety of search interfaces, web page formats, and API's. To complicate matters, some databases periodically change their export formats, effectively breaking the tools that provide access to their data. Although this scenario is an improvement on a decade ago, there is still very little Semantic Web technology involved. Most -omics databases do not yet provide metadata and, when it is available, do not provide it in a standard format with common semantics. We envision a future where not only data but also schemas that describe them are accessible in semantic web formats such as

RDF, RDFS, and OWL, and data is provided with the semantic annotations that are necessary to link them to the concepts that describe their components.

One of the most promising notions that Semantic Web brings to biology is that the search for data, and even experiments themselves, can eventually be specified in terms of the relevant biological concepts. Once disclosed along with the corresponding data, these concepts can then serve as part of the documentation for the experiment itself, helping to encode the hypothesis and relevant domain knowledge. Moreover, these concepts can eventually be used to define and steer the execution of a computational experiment, thus removing the burden of many implementation details and allowing scientists to define experiments in their own terms, i.e. ontologies of their choice or making.

An essential element of a semantic web contribution for the life sciences is data integration. Most forms of computational biology, workflow, data analysis, and visualization require data integration (see [1, 2] and references therein). In the context of the Virtual Laboratory e-science (VL-e) project, we consider how biologists could perform integrative bioinformatics research by considering biology use cases as working examples. Our specific use case requires data integration in order to explore the viability of a hypothesis that links epigenetics and transcription. Our goal is to perform data integration in a way that is repeatable and self-documenting as a result of syntactic and semantic disclosure. In this article, we describe a semantic web approach to performing biological data integration that we think is general enough to be useful to a variety of disciplines in the context of virtual laboratories and e-science.

Biology use case - Background

The goal of our use case is to unravel the relationship between the histone code, DNA sequence, and transcription. We start our incremental approach by studying the relationship between two components: histones and transcription factor binding sites in the DNA sequence. Histones are specific types of proteins that bind DNA and as such are central to packaging long DNA molecules into chromosomes in the nucleus of a cell. They undergo specific modifications, such that a pattern over the chromosomes is formed, referred to as a 'histone code' [3]. 'Transcription factors' are also proteins that can bind DNA to directly influence gene expression. Many of the DNA sequences to which transcription factors bind, i.e. transcription factor binding sites, have been identified and localized on human DNA. The biological question in our case is: How is chromatin involved with transcription?

Creating application-specific ontologies for *myModel*

The first step of our approach is to assemble the concepts relevant to our biological hypothesis. These concepts will serve as the terms of a controlled vocabulary that we can use to build our queries. The use of ontologies as controlled vocabularies can be found in practice such as in [4]. Lacking an existing ontology that covers the concepts

relevant to our research problem, we create an ontology that will serve as *myModel*¹. This ontology is limited to the concepts necessary to describe the problem domain of our experiment and could be called an *application-specific ontology*. We expect our ontology to evolve during the course of our case studies, or be merged with ontologies created by the domain community. These same OWL classes can be eventually used to link to other knowledge, such as relations with other OWL classes or rules [5] for use by reasoners. Although the ontology is used as a namespace during our data integration, the consistency checking that is possible is an important advantage of using OWL for *myModel*. Of course, best practices and design patterns [6] should be employed to ensure correctness and reusability. To enhance future interoperability with biomedical ontologies, we are investigating how to employ the relations proposed by Smith et al [7].

After evaluating the Gene Ontology [8] and seeking other appropriate ontologies using Swoogle [9], we decided to build our own application-specific ontology for histones in OWL: *HistOn*. One of the main reasons for doing this was to include a level of detail related to histones that we did not encounter in existing ontologies. We used the OWL plug-in of Protégé [10, 11]. To facilitate future reuse of the major parts of *HistOn* we created separate OWL files for each (the combined ontology can be viewed at [12]):

myModel consists of the following ontologies:

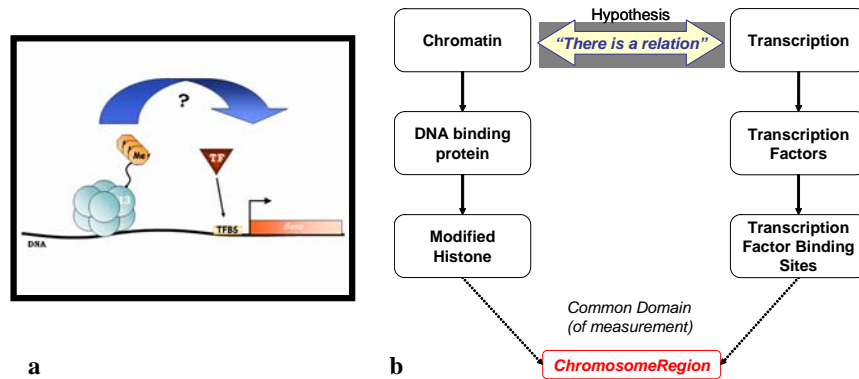
- Higher-level concepts related to epigenetics², such as *ChromosomeRegion*
- Histones and concepts directly related to histones
- Transcription factor binding sites and directly related concepts

Asserting the hypothesis

To represent the hypothesis for our use case, we started by drawing a cartoon, in line with common practice in the field of biology ([13]; Figure 1a). It shows, in ‘biologist-readable’ form, that we want to study the relationship between a particular histone (H3) with a particular modification (tri-methylation on the fourth lysine, i.e. H3K4Me3), and transcription factor binding sites (TFBS) because they are related to gene expression, and that both these elements are located on the DNA of a chromosome. In contrast to common practice in biology, we added concepts from the cartoon to *myModel*. Once we have added these concepts to *myModel* we can define the hypothesis in terms of this model (Figure 1b). In our example, the hypothesis is simply that there is a relation between the two concepts of chromatin and transcription. Knowledge representation such as that for hypothetical assertions remains future work.

¹ We use “*model*” to mean the machine-readable qualitative model relevant to the phenomenon being studied.

² Epigenetics refers to the heritable control over gene expression that is not linked to the DNA sequence alone; histones are likely to play an important role in this control.



Figures 1 (a) Cartoon representation (b) Schematic overview of the path to a common domain

Finding relevant data

Lacking a data broker or semantic mediator (see *Related Work*), most biologists must rely on their incidental knowledge of web resources to find appropriate data sources. In our case, this led us to the UCSC genome browser [14, 15]. The curators use an annotation strategy based on the genomic alignment algorithm BLAT [16]. In principle, any biological entity that can be associated with a DNA sequence can be localized on the chromosomes (human, rat, and mouse are among the species available at UCSC). Data from the ENCODE project, including histone binding data, is stored here, as well as transcription factor binding sites (TFBS).

Data import

We would like our final RDF data to contain information about original table structure, data source, the entry or row, syntactic type, and semantic type. We separate the import process into two steps: a syntactic and a semantic annotation step.

The syntactic step makes use of table information provided by UCSC. We defined an OWL schema (*theirDataModel*) to represent the table structures, both for the histone data and the transcription factor binding sites. This part of the import is similar to that used in YeastHub [17], where table column names are translated directly into RDF types, resulting in an RDF version of the UCSC table structure for the data. Each row of the table corresponds to a single measurement and requires a unique identifier, which we generate based on source file and row number. In this particular use case, we are not confronted with the more difficult but related problem of using a globally unique identifier (e.g. for a gene) such as is being discussed in the research community (see for example related discussions in [18]). Our measurement identifier is only required to be unique over the set of measurements that are being integrated. We also used the type information provided by UCSC in the form of a MySQL dump to generate the appropriate XML Schema Definition (XSD) tag for

each piece of data³. The XSD tag was to ensure that our data was properly interpreted (for example, not as a string in a comparison meant for numbers). The translation of the UCSC data into RDF was performed using a version of Mapper [19] that we modified for RDF output [20].

The semantic annotation step of the import process requires a mapping from the RDF types in the data (*theirDataModel*) to the corresponding semantic types in *myModel* (a model from the data producers, *theirModel*, has not been supplied). The purpose of the mapping is to enable our “semantic query” to be made in terms of *myModel* so we want to create the conditions where our query selects data with *theirDataModel* types when the query actually contains our own *myModel* types. We found that a subclassing of a *theirDataModel* property to a *myModel* property with `rdfs:subPropertyOf` produced the desired effect, both for that property and the RDF nodes at its endpoints (due to the RDFS reasoning in Sesame). This type of subtype mapping should also be possible in the case that a data provider supplies semantic metadata (i.e. *theirModel*) although the required ‘ontology alignment’ could be more complex⁴. Note that we could have directly translated column headers into the corresponding equivalent OWL type during the Mapper to RDF step and make RDFS reasoning unnecessary. However, such an approach would shift control of the mapping from the RDFS statements to the Mapper import stage and subsequent changes to *myModel* could require building an entire new RDF graph of the data with the new names that have resulted from the changes.

A basis for comparison: Finding the common domain for integration

In order to integrate measurement data, we must align values along the same domain or axis. We will take our use case as an example. We can use the graph of our ontology to look for such a domain. The comparable domain in our use case is the region defined by the class ‘*ChromosomeRegion*’. A chromosome region is an interval of DNA sequence located along a particular chromosome. In terms of the concept graph, *ChromosomeRegion* forms the link between the two concepts that we want to compare: both histones and transcription factor binding sites are related to this concept.

We also need to establish the criteria that make a given pair of measurements comparable, i.e. the measurements should be sampled from the same part of the domain. When there is overlap in the measurement domain, we want to compare the measurement values from the two different data sources corresponding to the overlap. To begin with, we chose a simple overlap criterion for our *ChromosomeRegion* intervals that we could encode directly in an RDF query.

³ We ran into a technical problem for large numbers (> 6M) of XSD tags and describe the solution at <http://integrativebioinformatics.nl/histone/HistoneDataIntegration.html> .

⁴ Ontology alignment is an area of ongoing research.

Data integration query

Once we have found a way to determine which measurement data can be meaningfully paired, we can perform the final step of our data integration experiment. Although it is possible to write a program to achieve this step, we chose to write an RDF query (see [21] for details). In this way, the semantics of the integration are readily available for inspection: all terms used in the query refer to OWL classes. With RDFS reasoning on, we can take advantage of the subsumption equivalence to our own myModel names, i.e. write the query with our own OWL “terms”. Our query returns a list of data pairs that can then be further explored by browsing, visualization, statistics (e.g. correlation), and data mining. Our preliminary results suggest that a large number of TFBS types are preferentially located within the regions overlapping H3K4Me3 binding sites, which is in line with experiments that suggest a role for H3K4Me3 in gene activation [22]. Further biological characterization of these TFBSs is work in progress.

Performance issue for ‘interval join’

The type of query that we use is called an ‘interval join’ in temporal and multimedia databases, where regions of media are checked for overlap with the regions defined for corresponding annotations. The ‘interval join’ appears to be unavoidable when performing data integration of measurements by query. Our largest datafile (all data for the genome) contains approximately 11M triples. Scalability and performance issues arose during initial tests, forcing us to run with smaller data and try different configurations. In this phase, we did not use RDFS reasoning and started performing queries in terms of *theirDataModel*. We created test data from a smaller data set (chromosome 22), and tried the query in several RDF systems⁵. It has been suggested [23] that different combinations of data and query can produce widely varying results. We indeed found that our query/data scaled unpredictably depending on the RDF implementation being used, with our query on the full data set taking on the order of days (see Table 1 and *Disclaimer*). In contrast, several non-RDF implementations executed the query in a matter of seconds. This discrepancy and significant performance differences between the RDF implementations themselves points to a performance bottleneck that could be better supported in RDF query engines, perhaps with custom support for our particular type of join. Note that in the case of MonetDB, custom optimizations for interval joins (called “StandOff joins” in [24]) can result in dramatic speedup in XQuery. However, although MySQL apparently performs better than RDF implementations with our query, the ‘interval join’ would still be too demanding for a public server based on a MySQL database⁶. Although space limitations prevent us from including the text of an actual query here, example queries can be found at [21].

⁵ Although the terms of our license agreement do not allow us to publish the performance results in our table at this time, our tests with the RDF implementation of an anonymous major DB vendor produced no improvements.

⁶ This could be done with the translation from RDF in the case of a query rewriting approach such as that employed for D2RQ [34].

Table 1: Naïve comparison (unequal platforms and technologies)

	Chromosome 22	Genome
SWI Prolog (RDF)	3.25 minutes	X
Sesame (RDF)	4.25 minutes	42 hours
Jena (RDF)	8 minutes	8 days
Python custom program	X	17 seconds
XQuery (MonetDB)	X	7 seconds
mySQL	X	98 minutes

Disclaimer: This table is a naïve comparison and not a benchmark! Different conditions exist between tests e.g. machines, machine load, level of configuration and API expertise, version numbers, etc.

Related Work

Data integration is an important topic in biology, and numerous solutions have been developed to enable retrieval of data from heterogeneous distributed sources (for reviews see e.g. [1, 2, 25]). The solutions range from monolithic, such as SRS [26] that uses keyword indexing and hyperlinks, Kleisli/K2 [27] that uses a query language that can query across databases as if they were one, data warehouses such as BioZon [28], to solutions that use web services acting as portals to biological data [29]. Perhaps the most widely used system is SRS, providing integration of more than 400 databases.

Our approach to data integration uses semantic models to provide a schema for integration. TAMBIS pioneered such an approach by creating a molecular biology ontology as a global schema for transparent access to a number of sources including Swiss-Prot and Blast [30]. Systems such as BACIIS [31], BioMediator [32] and INDUS [33] extend on this example. For instance, BioMediator uses a ‘source knowledge base’ that represents a ‘semantic web’ of sources linked by typed objects. The knowledge base includes a ‘mediated schema’ that can represent a user’s domain of discourse. INDUS shows important similarities to our approach. INDUS offers an integrated user interface to import or create user-ontologies (similar to ‘myModel’, but limited to ‘attribute-value hierarchies’), and create ontological mappings between concepts and ‘ontology-extended’ data sources. In contrast to our approach, however, INDUS does not use semantic web formats such as OWL and RDF. While the syntactic step of our import is similar to that of YeastHub [17], our explicit linking of the semantic types to the syntactic types with RDFS moves the work of discovering the semantics from the query stage to the stage of model alignment.

Future Work

Our import process is meant to eventually create a transparent data access layer (termed *wrapper* in BACIIS) to external data sources such as UCSC. The import approach that we have described here is being extended to work on all UCSC data.

The translation to RDF can be automated with the use of the MySQL table information provided by UCSC. The table information can be used to create both the *theirDataModel* in OWL and the XML “map” used by the Mapper program for a standardized mapping from UCSC data to RDF. Of course, our approach is general and can eventually be applied to other data providers. However, we expect that data providers such as UCSC will add RDF export to their set of services. Once RDF export is available, a general data mediation (web) service becomes possible. Approaches that make use of a mapping between RDF and relational database schemas such as D2RQ [34] could eventually be used to provide data access via RDF queries.

Our query results in a set of abstract *overlaps* or derived features, i.e. regions of a domain from which measurements have been taken that we have deemed ‘interesting’ according to a criterion (in our case, overlap in the domain of measurement). These *overlaps* contain information relevant to our hypothesis. A challenge for computational experimentation is to create an OWL class for use in the semantic annotation of these *overlaps* that exposes them to queries for data related to *Chromatin* and *Transcription*.

Discussion

We propose a semantic web approach to data integration and report on our experience applying it in the context of a biological use case. Our approach is model-oriented and allows us to perform data integration experiments in terms of our own biological knowledge. It allows us to perform data integration of experimental measurement with a query in terms of *myModel*. This type of ‘semantic disclosure’ exposes meaning and application logic that would otherwise only be available to scientists that can interpret the code of the application that uses the data.

It appears that an interval join is unavoidable wherever measurement data is to be integrated with a query language. Better support for interval joins in RDF query engines are therefore important for adoption of this approach for exploratory analysis, where interactivity is generally preferred.

One of the semantic web bottlenecks that we have encountered is the general lack of semantic disclosure: i.e. *theirModel* (semantic model) is not supplied by the data provider. Such information is especially crucial to efforts that attempt to facilitate data integration that crosses domains of expertise. A more practical reason is that in some cases it is difficult to discover what the biological data really means. Therefore, we find it encouraging that *theirModel* could become available from, for example, NCBI [35] in the future.

Acknowledgements

We thank Willem van Hage for his assistance with RDFS features of Sesame. Thanks to Peter Boncz, Bart Heupers, Jacco van Ossenbruggen, and Jan Wielemaker (as well as colleagues at the anonymous major DB vendor) for help with the tests in Table 1.

This work was carried out in the context of the Virtual Laboratory for e-Science project (<http://www.vl-e.nl>), and the BioRange program of the Netherlands Bioinformatics Centre (NBIC). VL-e is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ). BioRange is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

References

- 1 Searls DB. Data integration: challenges for drug discovery. *Nat Rev Drug Discov* 2005; 4(1): 45-58.
- 2 Stein LD. Integrating biological databases. *Nat Rev Genet* 2003; 4(5): 337-45.
- 3 Strahl BD and Allis CD. The language of covalent histone modifications. *Nature* 2000; 403(6765): 41-5.
- 4 **About BIRNLex** [<http://xwiki.nbirn.net:8080/xwiki/bin/view/BIRN-OTF/About+BIRNLex>]
- 5 **Rule Interchange Format Working Group Charter** [<http://www.w3.org/2005/rules/wg/charter>]
- 6 **SWBP&D WG Semantic Web Tutorials** [<http://www.w3.org/2001/sw/BestPractices/Tutorials>]
- 7 Smith B, et al. Relations in biomedical ontologies. *Genome Biol* 2005; 6(5): R46.
- 8 Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; 25(1): 25-9.
- 9 Ding L, et al. Swoogle: a search and metadata engine for the semantic web. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM Press: Washington, D.C., USA. 2004. 652-659.
- 10 **Protégé** [<http://protege.stanford.edu/>]
- 11 Knublauch H, Dameron O, and Musen MA. Weaving the Biomedical Semantic Web with the Protégé OWL Plugin. *First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004)* 2004 (Whistler (BC, Canada)), American Medical Informatics Association; 33-47.
- 12 **OWLDocs of Overview Ontology for myModel** [<http://integrativebioinformatics.nl/histone/OWLDocs/OverviewOntology/index.html>]
- 13 Perini L. Explanation in Two Dimensions: Diagrams and Biological Explanation. *Biology and Philosophy* 2005; 20: 257-269.
- 14 Gribskov M. Challenges in data management for functional genomics. *Omics* 2003; 7(1): 3-5.
- 15 Kent WJ, et al. The human genome browser at UCSC. *Genome Res* 2002; 12(6): 996-1006.

- 16 Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* 2002; 12(4):
656-64.
- 17 Cheung KH, et al. YeastHub: a semantic web use case for integrating data in
the life sciences domain. *Bioinformatics* 2005; 21 Suppl 1: i85-i96.
- 18 **Semantic Web for the life sciences discussion forum**
[\[http://lists.w3.org/Archives/Public/public-semweb-lifesci/\]](http://lists.w3.org/Archives/Public/public-semweb-lifesci/)
- 19 **Navigate data with the Mapper framework, Build your own data
mapping system with an interlingual approach**
[\[http://www.javaworld.com/javaworld/jw-04-2002/jw-0426-mapper.html\]](http://www.javaworld.com/javaworld/jw-04-2002/jw-0426-mapper.html)
- 20 **Mapper** [https://gforge.vl-e.nl/projects/mapper/]
- 21 **Semantic Data Integration for Histone Use Case Website**
[\[http://integrativebioinformatics.nl/semanticdataintegration.html\]](http://integrativebioinformatics.nl/semanticdataintegration.html)
- 22 Schubeler D, et al. The histone modification pattern of active genes revealed
through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev*
2004; 18(11): 1263-71.
- 23 **Pitfalls in Benchmarking Triple Stores**
[\[http://jeenbroekstra.blogspot.com/2006/02/pitfalls-in-benchmarking-triple-stores.html\]](http://jeenbroekstra.blogspot.com/2006/02/pitfalls-in-benchmarking-triple-stores.html)
- 24 Alink W, et al. Efficient XQuery Support for Stand-Off Annotation.
*Proceedings of International Workshop on XQuery Implementation,
Experience and Perspectives (XIME-P)* 2006 (Chicago, IL, USA).
- 25 Eckman B, Rice J, and Schwarz P. Data management in molecular and cell
biology: vision and recommendations. *Omic* 2003; 7(1): 93-7.
- 26 Zdobnov EM, et al. The EBI SRS server-new features. *Bioinformatics* 2002;
18(8): 1149-50.
- 27 Ritter O, et al. Prototype implementation of the integrated genomic database.
Comput Biomed Res 1994; 27(2): 97-115.
- 28 Birkland A and Yona G. BIOZON: a hub of heterogeneous biological data.
Nucleic Acids Res 2006; 34(Database issue): D235-42.
- 29 Wilkinson M, et al. BioMOBY successfully integrates distributed
heterogeneous bioinformatics Web Services. The PlaNet exemplar case.
Plant Physiol 2005; 138(1): 5-17.
- 30 Stevens RD, Robinson AJ, and Goble CA. myGrid: personalised
bioinformatics on the information grid. *Bioinformatics* 2003; 19 Suppl 1:
i302-4.
- 31 Ben Miled Z, et al. An efficient implementation of a drug candidate
database. *J Chem Inf Comput Sci* 2003; 43(1): 25-35.
- 32 Mork P, Shaker R, and Tarczy-Hornoch P. The Multiple Roles of Ontologies
in the BioMediator Data Integration System. *DILS* 2005, Springer; 96-104.
- 33 Caragea D, et al. Algorithms and Software for Collaborative Discovery from
Autonomous, Semantically Heterogeneous, Distributed Information Sources.
ALT 2005, Springer; 13-44.
- 34 **D2RQ** [<http://www.wiwiss.fu-berlin.de/suhl/bizer/d2rq/spec/>]
- 35 **public-semweb-lifesci forum message from Benjamin H. Szekely**
[\[http://www.w3.org/mid/OFC5D7E901.5F3825EB-ON85257169.0060CA27-85257169.006B0FEE@us.ibm.com\]](http://www.w3.org/mid/OFC5D7E901.5F3825EB-ON85257169.0060CA27-85257169.006B0FEE@us.ibm.com)