Journal of Biomedical Informatics 41 (2008) 706–716

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Bio2RDF: Towards a mashup to build bioinformatics knowledge systems

François Belleau [a,*], Marc-Alexandre Nolin [a,b,*], Nicole Tourigny [b], Philippe Rigault [a], Jean Morissette [a,c]

[a] Centre de Recherche du CHUL, Université Laval, 2705 Boulevard Laurier, Que., Canada G1V 4G2
[b] Département d'informatique et de génie logiciel, Université Laval, Cité Universitaire, Que., Canada G1K 7P4
[c] Département d'anatomie-physiologie, Université Laval, Cité Universitaire, Que., Canada GIK 7P4

ARTICLE INFO

Article history:
Received 1 September 2007
Available online 21 March 2008

Keywords:
Knowledge integration
Bioinformatics database
Semantic web
Mashup
Ontology

ABSTRACT

Presently, there are numerous bioinformatics databases available on different websites. Although RDF was proposed as a standard format for the web, these databases are still available in various formats. With the increasing popularity of the semantic web technologies and the ever growing number of databases in bioinformatics, there is a pressing need to develop mashup systems to help the process of bioinformatics knowledge integration. Bio2RDF is such a system, built from rdfizer programs written in JSP, the Sesame open source triplestore technology and an OWL ontology. With Bio2RDF, documents from public bioinformatics databases such as Kegg, PDB, MGI, HGNC and several of NCBI's databases can now be made available in RDF format through a unique URL in the form of http://bio2rdf.org/namespace:id. The Bio2RDF project has successfully applied the semantic web technology to publicly available databases by creating a knowledge space of RDF documents linked together with normalized URIs and sharing a common ontology. Bio2RDF is based on a three-step approach to build mashups of bioinformatics data. The present article details this new approach and illustrates the building of a mashup used to explore the implication of four transcription factor genes in Parkinson's disease. The Bio2RDF repository can be queried at http://bio2rdf.org.

1. Introduction

A rapid way to look for information on the web is to use a search engine such as Google. The results, however, are a list of suggested HTML pages devoid of context and semantics and requiring human interpretation. For a more contextual search in the field of molecular biology, a specialized tool like NCBI's Entrez [1] is more effective because it is dedicated to the specific domain under consideration. The Entrez search engine uses all the different databases hosted by NCBI; its data integration approach, based on hyperlinks, is illustrated by its database schema (http://www.ncbi.nlm.nih.gov/Database/). The Kegg's DBGET [2] search service is another example of a specialized search engine dedicated to genes and pathways.

Each year, NAR [3] publishes a new version of its bioinformatics database list. In the 2006 issue, over one thousand servers were listed. Other specialized lists of databases are now available. For instance, the Pathguide website [4] lists 244 pathways and protein interaction databases. With such a proliferation of knowledge sources, there is a pressing need for a global multisite search engine and for good data integration tools. According to the data warehouse approach, such services can be built by collecting information into a central data repository [5] and queried with an interface built on top of the repository. However, the warehouse approach does not address the problem of accessing a database outside the warehouse. A system that would be able to query and connect different databases available on Internet would solve that problem. This is one of the goals of the semantic web approach: to offer the data warehouse experience without the need of moving first the data into a central repository.

To address the data integration problem, the semantic web community, led by the W3C, proposed a solution based on a series of standards: the RDF format for document [6] and the OWL language for ontology specification [7]. RDF and OWL generate a series of entities called 'triple' in the form of a subject, predicate and object. Database systems able to handle triples are called triplestore. New software has been created by the computer science community to exploit them. Some tools are still in the development stage, others are mature enough to be used in production systems, like the open source project Sesame [8], which is a triplestore server providing storage and querying capabilities.

We have developed a semantic web application called Bio2RDF to help solve the problem of knowledge integration in bioinformatics. Bio2RDF uses RDF documents and a list of rules to create URIs

* Corresponding authors. Address: Département d'informatique et de génie logiciel, Université Laval, Cité Universitaire, Que., Canada G1K 7P4. Fax: +1 418 525 4444x42761 (M.-A. Nolin).
E-mail addresses: francoisbelleau@yahoo.ca (F. Belleau), Marc-Alexandre.Nolin@genome.ulaval.ca, lotus@ieee.org (M.-A. Nolin).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.03.004

that will create linked data. Bio2RDF can be seen as a mashup application because it combines data from more than one source, following the definition of a mashup given in Wikipedia [9]. Indeed, Bio2RDF integrates publicly available data from some of the most popular databases in bioinformatics. As a mashup is more often associated with a graphical user interface than data (or knowledge) integration, Bio2RDF can be described as a data mashup using a semantic web approach for data (or knowledge) integration. The purpose of the present paper is to describe the data integration approach used with Bio2RDF.

1.1. Integration methods in bioinformatics

The idea of integrating data from various sources is not a recent concern in bioinformatics, as illustrated by the research work of Davidson [10], Köhler [11], and Stein [5].

In 1995, Davidson [10] suggested the following basic steps to integrate bioinformatics data: transformation to a common data model, matching of semantically related objects, schema integration, transformation of data into a federated database, and finally matching of semantically equivalent data. Davidson et al. suggested to "Transform data to the federated database on demand". This solution can now be achieved in a semantic web approach through the Bio2RDF project, where data is transformed into RDF format.

In 2003, the Semeda (Semantic Meta Database) [11] was another attempt at integrating heterogeneous databases. Kohler identified four problems. (1) In different databases the same things can be given different names. This is the case with the two pathway databases, Kegg [12] and Reactome [13]: they both annotate and describe the same pathways in completely different semantic spaces. (2) Attribute names are not self-explanatory. For example the way of specifying URLs should always be the same, as in the HTML href attribute. (3) Querying databases requires knowledge about its contents. This is exactly what the semantic web approach wants to avoid. (4) Due to the lack of a systematic linking mechanism, only the most important attributes are associated. Therefore, a normalization of identifiers is mandatory. Such a normalization was the goal of the LSID [14] project.

Also in 2003, Stein [5] highlighted three approaches typically used by data integrators: link integration, view integration, and data warehousing. The first one uses the linking capability of the web; the second one is the creation of portals that aggregate the information; the third, data warehousing, stores everything in a single unified database. Stein also proposed an ontological approach that he called knuckles-and-nodes. Simply stated, this approach is about building databases of links between data, but not storing any of it. This strategy is very similar to that of Bio2RDF.

1.2. Integration using a semantic approach

Ontology design is not a new topic in bioinformatics, however projects using the OWL language are new. Tambis [15], BioPAX [16] and UniProt [17] are three projects which have adopted this new formalism. Describing and building knowledge systems using the semantic web's RDF standard as a knowledge representation format is still a challenge and several projects such as YeastHub [18] and FungalWeb [19] have explored this research topic.

In 2000, TAMBIS [15] was the first project to propose a unified ontology described in OWL and covering many aspects of the bioinformatics knowledge space. The BioPAX ontology [16], a more recent proposition with the same goal, is already used by six pathway database websites. The UniProt consortium has made available an RDF version of the UniProt protein knowledge base through their new beta website (http://beta.uniprot.org). The documented translation [20], describing the migration from the UniProt traditional text format to an RDF document has been a guideline for the Bio2RDF project. Its ontology [21], available in OWL format, was created with the Protégé ontology editor [22].

The YeastHub [18] project was the first attempt to build an integrated database in RDF format unified by the Sesame's triplestore. The resulting warehouse of yeast genome data illustrates the potential of the query capabilities afforded by a knowledge base once the document's URIs have been normalized. The Bio2RDF approach is similar to that of YeastHub, with the exception that Bio2RDF is open source, extensible and provides access to millions of documents from hundreds of different organisms.

The FungalWeb [19] project also focused on data integration, specifically for the needs of industrial enzyme biotechnology. An instantiated OWL-DL ontology was designed using Protégé and the graphical query composer OntoIQ [23], in conjunction with Racer and its query language nRQL. The interrogation of the integrated knowledge base was illustrated by using application scenarios. Instead of using Sesame, this research project employed the commercial OWL reasoner Racer [24] which offers inference capabilities.

A third integration project using RDF, conducted by Stephens [25], integrated disparate biomedical data sources to help the drug discovery effort. Different data sources were merged together: UniProt, OMIM [26], Entrez Gene [27], Kegg, Gene Ontology [28], Intact, Affymetrix probesets annotations and some others. This list of major data sources is similar to that of Bio2RDF. To build this knowledge base system, Stephens used the Oracle RDF data model as the triplestore and the Seamarks Navigator for faceted browsing. Bio2RDF is also an integration project making bioinformatics data available on the web from various data sources, but uses open source software. This framework does not offer a user interface with faceted browsing, but tools like Simile Exhibit can be used directly with the Bio2RDF data.

In a review about data integration and genomic medicine [29], the authors have identified two axes defining the integration approach. The first one describes the architecture of the system, the second axis defines the knowledge description. Using this definition, Bio2RDF should be classified into Peer data management systems with an ontology knowledge description.

Several lessons were learned from these experiences. Firstly, the semantic web approach can be used effectively to integrate bioinformatics data. Secondly, knowledge bases created thus far were designed to answer specific questions. Thirdly, if one wants to promote the semantic web method for data integration, the use of free open source software should be encouraged in order to enhance the reproducibility of results that are published in the literature. The Bio2RDF project was built as a result of these lessons.

The present article intends to show how Bio2RDF merges bioinformatics knowledge from different sources. Aggregation of related knowledge sources should eventually be as easy as dragging and dropping them into a knowledge store. The Bio2RDF integration technology is built on programs found in the open source community: the Sesame triplestore and Elmo RDF crawler [8], JSP and JSTL [30] which are technologies used to generate web pages and the URLrewrite library [31] used to proxy HTTP requests. RDF-formatted documents, required by semantic web technologies, are not yet common on Internet. At this time, only UniProt and GO websites offer RDF documents to build semantic web applications. One of the main goals of the Bio2RDF project is to convert into RDF format documents available from public databases. Bio2RDF is a flexible open source software which allows to develop new rdfizer programs in order to add new knowledge sources or experimental private data. The result section below shows, through a use case, how the Bio2RDF mashup system can be used to build a triplestore that supports the exploration of the Parkinson's disease knowledge space.