Bio2RDF: Towards A Mashup To Build Bioinformatics Knowledge System

François Belleau[1,2*], Marc-Alexandre Nolin [1,2**],

Nicole Tourigny[2], Philippe Rigault[1], Jean Morissette[1]

[1]Centre de Recherche du CHUL, Université Laval
2705 Boulevard Laurier
Quebec, QC, Canada, G1V 4G2

[2]Département d'informatique et de génie logiciel,
Université Laval
Cité Universitaire Québec, QC, Canada, G1K 7P4

For correspondence:
*Francois.Belleau@genome.ulaval.ca, **Marc-Alexandre.Nolin@genome.ulaval.ca

ABSTRACT

With the increasing popularity of the semantic web technology and the ever-growing number of databases in bioinformatics, there is a pressing need to develop mashup systems to integrate bioinformatics knowledge. Bio2RDF is such a system, built from an rdfizer written in JSP, the Sesame open source triplestore technology and an OWL ontology. With Bio2RDF, all documents from public bioinformatics databases, GeneID, OMIM, UniProt, Kegg, Ligand, OBO, PDB and MGI are now made available to the scientific community in RDF format. A total of 140 gigabytes of XML and text data have been converted into 46 million RDF documents in the Bio2RDF repository. Each one is accessible through a unique URL in the form of http://bio2rdf.org/namespace:id or by an LSID. In addition to this service myBio2RDF is an open source application that lets the user drag and drop knowledge from the bioinformatics semantic web into a local Sesame triplestore where it can be further analyzed with SeRQL queries. The Bio2RDF project has successfully applied the use of 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. The Bio2RDF repository of RDF documents can be queried at http://bio2rdf.org. The myBio2RDF application, which is a modified version of Sesame triplestore with rdfizers, can be downloaded at http://sourceforge.net/projects/bio2rdf/. The project's OWL ontology is located at http://bio2rdf.org/bio2rdf-2007-02.owl

Categories and Subject Descriptors

E.1 [**DATA STRUCTURES**]: Distributed data structures, Graphs and networks

J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics

General Terms

Design, Standardization

Keywords

Knowledge integration, Bioinformatics database, Semantic web, Mashup, Ontology

WWW 2007, May 8--12, 2007, Banff, Canada.

1.INTRODUCTION

When searching for information, a rapid way to obtain a list of suggested HTML pages is to use a search engine such as Google. However, as good as the results can be, they are 'semanticless'. For a more relevant search, a specialized tool like NCBI's Entrez [27] is more effective because it is dedicated to the specific domain of molecular biology. This search engine parses all the different databases hosted by NCBI and its data integration approach, based on hyperlinks, is illustrated by its database schema (http://www.ncbi.nlm.nih.gov/Database/). Kegg's DBGET [11] search service is another example of a specialized search engine dedicated to genes and pathways.

Each year, NAR [10] publishes a new version of its bioinformatics database list and in the 2006 issue, over one thousand servers are listed. Others specialized lists of databases are now available. For example, the Pathguide website (http://www.pathguide.org/) lists 244 pathways and protein interaction databases. With such a proliferation of knowledge sources, there is a pressing need for a global multi-site search engine and good data integration tools. According to the data warehouse approach, such services can be built around a central repository of data where information is collected [29]. However, the actual warehouse approach does not address the problem of access to a database not yet integrated in a given data warehouse. A system that would be able to query different databases available on the Internet would solve that problem. Indeed, this is one goal of the Semantic Web: to offer the data warehouse experience without moving the data into a central repository.

To address the integration problem, the semantic web community, lead by the W3C, proposed a solution based on a series of standards: the RDF format for document (http://www.w3.org/RDF/) and the OWL language for ontology specification (http://www.w3.org/2004/OWL/). RDF documents and OWL ontologies can be converted into 'triple' entities, in the form of a subject, predicate and object, for which a variety of tools developed by the computer science exist. Some tools are still in the developmental stage, others are mature enough to be used in production systems, like the open source project Sesame [1], which is a triplestore server providing storage and querying capabilities. The goal of the Bio2RDF project is to solve the problem of knowledge integration in biology by applying a semantic web approach. It integrates publicly available data from some of the most popular bioinformatics databases. This corresponds to a mashup, defined as a semantic web application that combines content from more than one source into an integrated experience [31]. In this paper, an approach to build mashup is presented.

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 [8], Köhler [24] and Stein [29] .

In 1995, Davidson [8] 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 to a federated database and finally the matching of semantically equivalent data. Davidson *et al.* proposed to "Transform data to the federated database on demand". This solution can now be achieved using a Semantic Web approach applied in the Bio2RDF project where data is transformed to the RDF format.

In 2003, the Semeda (Semantic Meta Database) [24] was another attempt dedicated to the integration of heterogeneous databases. Kohler *et al.* 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 [22] and Reactome [21]: 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, like HTML href attribute. 3) Querying databases requires knowledge about its contents. Which 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 associate together. Therefore, a project to normalize identification is mandatory, which is the goal of the LSID [19] project.

Also in 2003, Stein [29] 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 is the creation of portals that aggregate the information and 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.

Integration using a Semantic Approach

Ontology design is not a new topic in bioinformatics, however projects using the OWL language are new. Tambis [30], BioPAX [15] and UniProt [20], are three projects that 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 [7] and FungalWeb [28]) have explored this research topic.

In 2000, TAMBIS [30] was the first project to propose a unified ontology described in OWL [37] that covered many aspects of bioinformatics knowledge. The BioPAX ontology [15], a more recent proposition with the same goal, is already used by six pathway database websites. The SwissProt project has made available an RDF version of the UniProt [20] protein knowledge-base. The documented translation [38], describing the migration from the UniProt traditional text format to an RDF document has been an excellent guideline for the Bio2RDF project. It's ontology [39], which is available in OWL format, was created with the use of the Protégé ontology editor [33].

The YeastHub [7] project was the first attempt to build an integrated database in RDF format unified by Sesame's [1] triplestore. This approach was re-used in Bio2RDF. The resulting warehouse for 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 [7], with the exception that Bio2RDF is opensource, extensible and provides access to millions of documents from hundreds of different organisms.

The FungalWeb [28] project also focused on data integration, specifically for the needs of industrial enzyme biotechnology. An instantiated OWL-DL ontology was designed using Protégé and using a graphical query composer OntoIQ [40], in conjunction with Racer and it's query language nRQL, interrogation of the integrated knowledge-base was illustrated using application scenarios. Instead of using Sesame, this research project used the commercial OWL reasoner Racer [41], which offers inference capabilities.

The Bio2RDF project has learned three lessons from experiences with tools and databases. Firstly, the semantic Web approach can be used to integrate bioinformatics data. Secondly, present knowledge bases were designed to answer specific questions. Thirdly, if one wants to promote the semantic Web method for data integration, the use of free and open source software should be encouraged since this can enhance the reproducibility of results that are published in literature. Bio2RDF project has learned from all these lessons.

The goal of this article is 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. Bio2RDF integration technology is built on programs found in the open source community: Sesame triplestore [1] and Elmo RDF crawler [2], JSP and JSTL [42] which are technologies used to generate web pages and the URLrewrite library [43] used to proxy http requests. To leverage Semantic Web technology, RDF documents are necessary but they are not yet common on the Internet. At this time, only UniProt [20] and GO [3] websites offer RDF to build semantic web applications. One of the main goals of Bio2RDF project is to make documents from public databases available in RDF format. Bio2RDF is a flexible open source software, which allows development of new rdfizer programs to add new knowledge sources such as interesting web sites or experimental private data. The result section hereafter shows, with a use case, how the Bio2RDF mashup system can build a triplestore that supports the exploration of the Parkinson's disease knowledge space.

2.MATERIAL AND METHOD

Two main ideas have oriented our software development: the conversion of existing databases into RDF format by a process called "rdfizing" and the use of existing semantic web software to merge, query and visualize them. These software components are: Sesame [1] open source triplestore, Protégé ontology editor [33], Piggy Bank [18] semantic web browser plugin for FireFox and Welkin [26] RDF graph visualizer both developed by the MIT and, finally, the experimental LSID browser [19].

Firstly we describe the method used to build the ontology, then explain how to use rdfizer programs to transform existing documents to RDF format and, finally, how we normalized URIs. This section ends with a high level description of the system software architecture.

Ontology Design

An ontology can be defined as an explicit specification of a conceptualization that is an abstract and simplified view of the world that needs to be represented for some purpose [13]. For a

given knowledge-base or knowledge system, it means that a conceptual language should be used to define the objects to be represented and their relations. OWL is the conceptual language chosen by the Semantic Web community for ontology-based knowledge representation. Protégé-OWL is an editor that supports the OWL language within Protégé, which is an open source knowledge base framework. The ontology of Bio2RDF was designed with Protégé-OWL.

Since the main Bio2RDF goal is to convert into RDF format existing documents that are available on the web (GeneID description of Hk1 on NCBI web site for example), the first step is to analyze the existing HTML page to identify the predicates and relations describing the entities. The label of a field corresponds to its predicate, and the hyperlink corresponds to URI of the resource usually defined in another namespace like GI, GO, or PubMed. Using this approach we produced an OWL description, created with Protégé, from each selected HTML document. This step was been repeated for each namespace recognized by the current version of Bio2RDF: GO, OMIM, PDB, etc. For BioPAX [15] and UniProt this step was unnecessary because their OWL schema is already available. Finally, the global bio2rdf-2007-02.owl ontology description was built by merging the ontology file of each namespace.

Once the Bio2RDF ontology was created, the second step consisted of writing the necessary rdfizer programs in JSP in order to address two key objectives: 1) mapping between the data elements of the original document and the predicates in the RDF version, 2) normalization of URI resources according to the Bio2RDF's syntax. The programming of the rdfizers for more than twenty different namespaces was our main task.

The design of Bio2RDF's ontology was inspired by existing ontologies. For instance, rdf:type and rdfs:label were systematicaly used in each document. The label predicate always contains the name of the resource followed by a short form of its LSID enclosed with "[ ]". For example, rdfs:label of genid:15275 is "hexokinase 1 (Hk1) [geneid:15275]". Some common predicates from the Dublin Core project [30] were used: dc:title, dc:identifier, dc:created and dc:modified. We also used the FOAF [6] namespace to describe people and the BibTeX [23] one for literature reference. We had to create our own predicates in the bio2rdf namespace, the ones most used were bio2rdf:url, bio2rdf:urlImage, bio2rdf:lsid, bio2rdf:xRref, bio2rdf:name and bio2rdf:synonym. Definition of the semantics of these predicates is found in the Bio2RDF ontology file [44].

Rdfizer Programs

In an ideal world, all data would be available in RDF format with complete normalization of URI and, consequently, all Internet documents would automatically connect together. However, this is not yet the case. While complete access through web pages to HTML version of data already exists, RDF version are not there. The Bio2RDF project provides normalized RDF documents from several data sources. A JSP toolbox has been created to generate RDF files from locally stored databases or directly from HTML documents accessed via http request. JSP tools help to build rdfizers, which are programs that transform existing data into an RDF representation [45]. Several different sources of data can be rdfized such as relational databases, text files, XML documents, and HTML pages. For each type of knowledge source, JSP programs convert data from the original source to RDF format. These programs use XPath, regular expression or SQL query to extract knowledge from the original data. For example:

- **ncbi-omim2rdf.jsp**: converts XML representation of OMIM records provided by the NCBI efetch service into RDF;

- **ensembl-g2rdf.jsp**: Ensembl provides an access to its MySQL relational databases. This rdfizer queries using SQL commands and then transforms the answer into RDF;

- **prosite2rdf.jsp**: this program retrieves HTML pages and extracts the data to be converted by using regular expressions.

The format of the RDF document produced by the Bio2RDF rdfizer is not a definitive one. For this reason, rdfizer program source codes are made available for customization. Rdfizers can be quickly written and it will be easy to modify them if the authoritative data format changes.

URI Normalization

The availability of RDF documents is not sufficient. External references, expressed as URI, need to be normalized to allow proper connection of triples. For example, a PubMed reference with identifier 12728276 can be referenced with PMID:12728276, pubmed:12728276 or PubMed:12728276. For a knowledge agent, normalized representation of URI is mandatory to ensure functional connection between triples.

Rules to convert URI have been adopted for this reason. The major one is that a URI is uniquely attributed to a document that describes an object like an ontology term, a gene description or a protein annotation. Every URI has to be written in lowercase. Since the RDF repository is case sensitive, a simple uppercase letter present in URI is sufficient to create another triple link and the connection does not occur. The syntax of a normalized URI is described by the following pattern:

```
http://bio2rdf.org/<namespace>:<identifier>
```

For example, the identification for the article 12728276 from PubMed would be written as http://bio2rdf.org/pubmed:12728276. Our obvious URI design rule states that document URI corresponds to the unique URL, which returns the document in RDF format from Bio2RDF.org server. Consequently, when a new document URI is added into the triplestore, triples, which refer to this new document, connect to it. Bio2RDF also offers a service to fetch the original document in XML format, from the data provider. This service is for programmers who want to develop their own rdfizer. The pattern in this case is:

```
http://bio2rdf.org/data/<namespace>:<identifier>
```

For example, http://bio2rdf.org/geneid:15275 returns the RDF version of the HK1 gene of mouse from NCBI's GeneID and http://bio2rdf.org/data/geneid:15275 returns the XML version of the original document that could be obtained with NCBI Efetch request.

Bio2RDF Architecture

Figure 1 presents a schematic description of Bio2RDF architecture. All the external data sources, in different formats (XML, Text, ASN.1, KGML and RDF), are listed on the left. These sources are processed in two different ways. The top group is composet of websites (Ligand [22], Kegg [22], GeneID [25], etc.) from which the entire database was downloaded to the Bio2RDF.org server. RDF documents from these sources are then accessible at high speed since they are obtained directly from the Bio2RDF.org server. Data from these websites are stored in a 141 Gb size MySQL database. Only documents requested from websites of the bottom group (Reactome [21], Prosite [17], PubMed, etc.) are rdfized directly from the original source. In

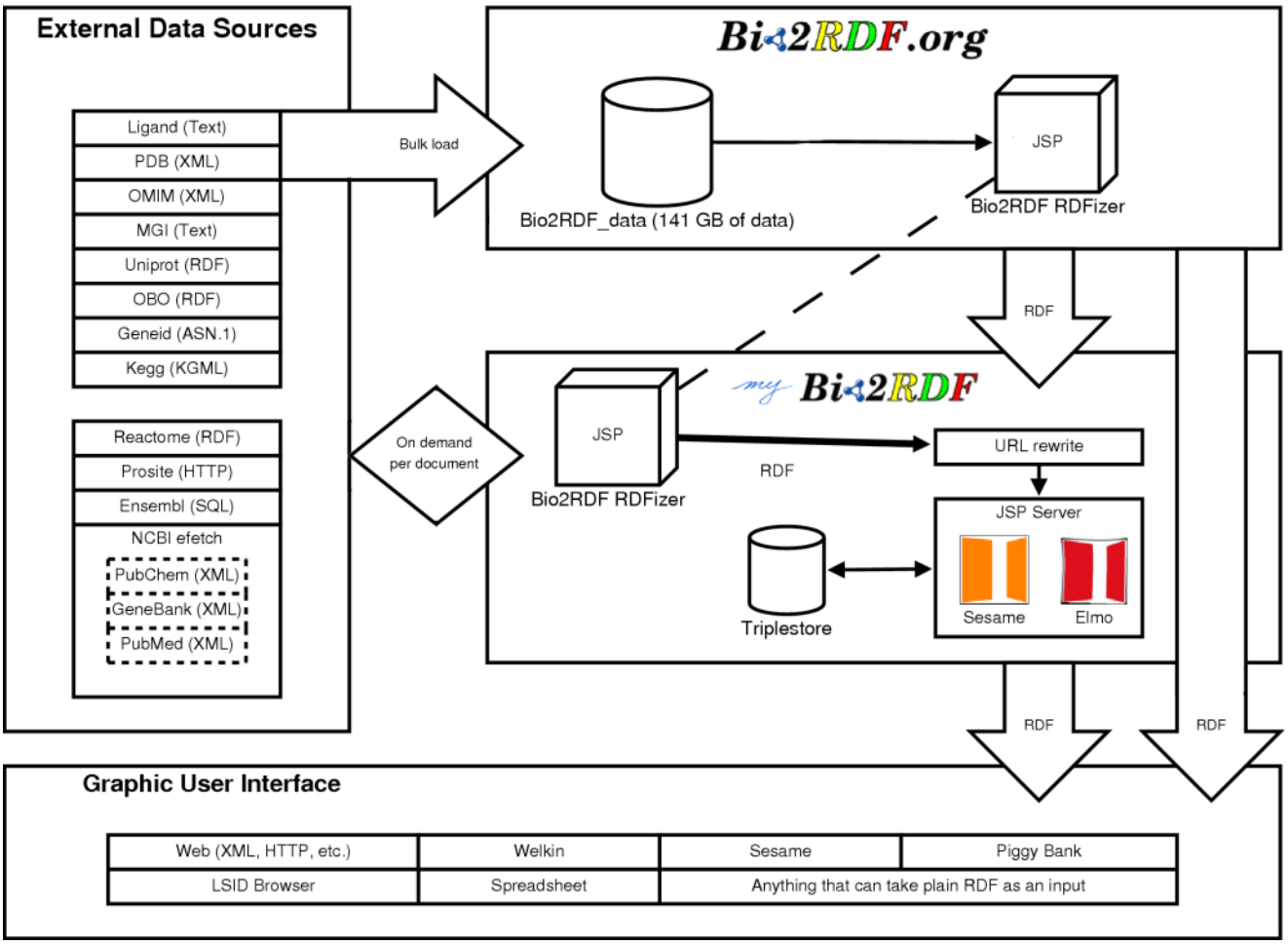

Figure 1: Bio2RDF knowledge system framework architecture

fact, the local rdfizer program, part of the myBio2RDF application, queries the data provider, transforms the returned document into normalized RDF and, finally, makes it available to the application. The data is cached for availability and speed purposes. Availability because few data providers have an RDF version of their documents and speed, because some data providers have restrictions on document access. Ultimately, the data should always come live from the data providers.

The myBio2RDF application contains two servlets running under a Tomcat server: Elmo [2] and Sesame [1]. Elmo is a RDF crawler which was originally created to follow `rdfs:seeAlso` predicate included in FOAF files. The Elmo capacity to crawl RDF documents from the Bio2RDF.org website, is applied to instantiate triples into a local Sesame repository where the requested documents are gathered. Next, the Sesame interface allows users to browse and query the knowledge-base with SeRQL. The Sesame version distributed with the myBio2RDF package was slightly modified to fit three special needs: 1) to allow its Explorer page to navigate through the external link defined with bio2rdf:url; 2) to see images defined by bio2rdf:urlImage; 3) to export query results in tabular format compatible with spreadsheets. The following services were added to allow Elmo to crawl specific knowledge :

`http://localhost:8080/search:TEXT@database`

`where database = [omim|geneid|pubmed|mesh|kegg|uniprot]`

- Obtaining a list of URIs corresponding to the results of a text search using the search engine of the corresponding web site.

`http://localhost:8080/load:NAMESPACE`

- Requesting all URIs in the triplestore which belongs to the specified namespace.

`http://localhost:8080/learn:NAMESPACE`

- Same as load, but will not reload already known node.

`http://localhost:8080/sameas:NAMESPACE1-NAMESPACE2`

- Creating a synonym node to link two URIs which have the same id but different synonymous namespace.

The URLRewrite library package matches the URL syntax with the appropriate RDFizer JSP program.

The software component controls the information workflow by interpreting the rules defined by regular expressions. Here is a sample of rules stored in URLRewrite configuration file. The first rule below calls ncbi-pubmed2rdf.jsp, a program that invokes the NCBI efetch utility to obtain the corresponding PubMed document in XML format and transforms it into RDF using XPath queries.

Original URL:

```
http://bio2rdf.org/jsp-bio2rdf/ncbi-
pubmed2rdf.jsp?id=12728276
```

The rule:

```
<rule><from>^/pubmed:(.*)</from>
<to>/jsp-bio2rdf/ncbi-pubmed2rdf.jsp?id=$1</to> </rule>
```

Resulting URL:

```
http://bio2rdf.org/pubmed:12728276
```

The next rule forwards to Bio2RDF.org server any URI request that cannot be locally resolved since there is no rdfizer program associated to the URL namespace. This forwarding rule chains Bio2RDF resolvers in the same way that DNS servers do.

```
<rule><from>^/(.*):(.*)</from>
<to type="redirect">http://bio2rdf.org/$1:$2</to></rule>
```

Since the Bio2RDF approach is flexible, it is possible to replace an existing rdfizer by another one simply by modifying the corresponding URL rewrite rule. For example, it is possible to write a rdfizer program to work with private data locally stored in a relational database. Once a new extension is added, new knowledge sources can then merge with Bio2RDF's existing ones. This is the way that the myBio2RDF application learns how to explore new knowledge space.

3.RESULTS

The Bio2RDF project is still in development, but some results are already available to the scientific community. More than twenty different public bioinformatics data sources are now available in normalized RDF format from the Bio2RDF.org server. This is a knowledge space of millions of documents. Some of the public databases were downloaded into a MySQL database and the rdfizer programs then converted documents.

Table 1: Number and size of RDF documents from public databases and available on the Bio2RDF server

Data source	URI example	Number of RDF documents	Size of data converted
go	go:0000001	22,961	507,963,321
kegg	path:aae00010	35,257	1,038,593,137
kegg	cpd:c00001	14,292	8,902,205
kegg	dr:d00001	4,153	2,985,017
kegg	ec:1.-.-.-	4,102	21,341,732
kegg	gl:g00001	10,951	5,259,938
kegg	rn:r00001	6,813	6,795,458
mgi	mgi:96103	70,175	210,458,897
ncbi	omim:100050	17,359	573,639,380
ncbi	geneid:1	2,744,786	67,225,535,082
obo	*obo's 59 namespaces*	279,720	216,007,267
pdb	pdb:100d	34,421	16,309,651,935
uniprot	uniparc:UPI0000000001	30,261,843	10,160,576,735
uniprot	uniprot:A0A000	4,177,176	29,453,203,064
uniprot	enzyme:1.-.-.-	5,020	2,844,058
uniprot	evoc:0100009	99	16,226
uniprot	keywords:1	892	567,252
uniprot	ontology:Apicoplast	10	5,388
uniprot	pubmed:100133	191,664	364,728,083
uniprot	taxonomy:10	337,564	125,630,659
uniprot	tissues:1	353	89,150
uniprot	uniref:UniRef100_A0A000	7,990,452	14,865,490,144
	Total	**46,210,063**	**141,100,284,128**

These databases are NCBI's GeneID [25] and OMIM [14], UniProt [20] protein knowledge base, Kegg's [22] pathway and Ligand [22] database, MGI [9] mouse annotations, OBO [35] open source ontology and PDB [4] the Protein Data Bank. The Elmo crawler [2] allowed downloading of RDF documents at high speed without stressing the official web site. By storing these major databases locally, hundreds of RDF documents, related to a specific topic, can be extracted in minutes rather than hours. Some other knowledge bases are also available in RDF format from the Bio2RDF server, although they are not hosted on it: Pubmed, GenBank, PubChem, Ensembl [16] and Reactome [21]. As an example, pathway definitions are offered in BioPAX [15] RDF format at the Reactome [21] website. These documents are accessed in real time from the usual website HTTP service. Table 1 gives the number of RDF documents downloaded from public databases and stored locally in our database. A total of 141 gigabytes of data was converted to RDF format. This knowledge space corresponds to 46 million of well-formed RDF documents using normalized URIs and respecting Bio2RDF ontology.

The downloaded databases had different formats. UniProt knowledge base was available in RDF format [20]. NCBI offered all its main databases in ASN.1 format from its ftp site where the GeneID [25] database was downloaded in ASN.1 format before its conversion to XML format. The OMIM [14] database was available only in tabulated text files. The efetch utility was used to individually extract each OMIM record in XML format. Gene Ontology [3] was available from three different sources (GO [3], OBO [35] and UniProt [20]) in three different RDF schemas. The GO's ftp server was chosen because it is the authoritative web site. The PDB [4] server releases all its records over an RSYNC server. Kegg's pathways [22] were downloaded from their ftp server in KGML (http://www.genome.jp/kegg/docs/xml/), an XML proprietary format. Compound, reaction and enzyme could only be downloaded in text format from the LIGAND database [22] and a Perl program was written to rdfize it. The MGI [9] mouse genome annotations, originally in tabulated text files, were transformed to RDF in the same manner. Finally, the OBO ontologies were downloaded in an RDF version. This format was previously produced by an experimental conversion project[46]. By using the Bio2RDF server or the myBio2RDF application, it is possible to browse millions of RDF documents using Sesame explorer in HTML page at http://bio2rdf.org/sesame, Piggy Bank [18] or the new experimental FireFox plug-in LSID browser [19]. The next section explains how an agent can automate this process and the role of Elmo [2] RDF crawler.

All statistics used in this article reflect the database conversions to Bio2RDF as of January 2007. Current statistics about Bio2RDF content are available at http://bio2rdf.org/namespace:all.

4.PARKINSON USE CASE

The potential of the Bio2RDF approach is illustrated by building a knowledge base about Parkinson's disease. This disease was chosen because it was already analyzed by the BioRDF subgroup of the HCLS community (http://www.w3.org/2001/sw/hcls/). A summary of concepts produced by a field expert is available at (http://esw.w3.org/topic/ParkinsonsDisease). At this stage, Bio2RDF is not yet a discovery system. It only accelerates the information gathering and may assist researchers in the information querying. In the following paragraph we show how the Bio2RDF mashup technology works and how a specific knowledge base about this disease can be built in 30 minutes by applying the myBio2RDF approach. This knowledge base can then be exploited to answer the following three questions:

– What is the semantic network of OMIM [14] records describing Parkinson's disease?

– Which MeSH terms are mostly cited in Parkinson's disease publications?

– What genes related to Parkinson's disease are involved in pathways according to Kegg [22] and Reactome [21] annotations?

Since we are interested in a genetic disease, we started to gather information from the OMIM website. Submitting the following URL to the Elmo crawler submits a research query to NCBI's Entrez search engine, which returns 166 OMIM identifiers. These URIs are then sent to Bio2RDF server to fetch the corresponding 166 documents in RDF format.

```
http://localhost:8080/bio2rdf/search:parkinson@omim
```

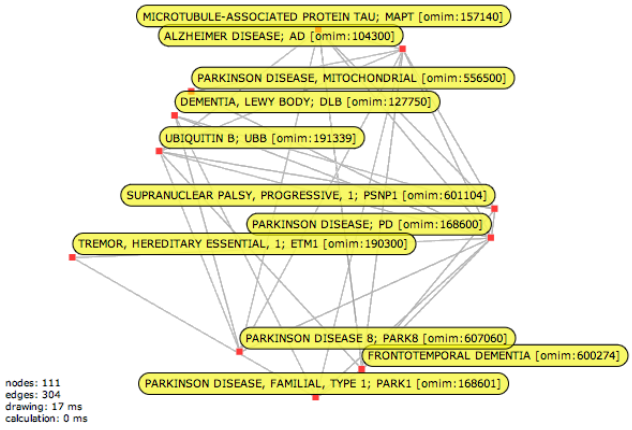

Figure 2: OMIM's graph of Parkinson's related records drawn with Welkin tool

This first operation created a local Sesame repository containing 166 OMIM records and the equivalent RDF file weighing 2.2 megabytes (Mb) and containing 14 kilo triples (Kt). Downloading these documents from Bio2RDF.org server took less than 2 minutes., This first SeRQL query was then submitted to the Sesame user interface that helps to visualize relations between the OMIM [14] records.

```
CONSTRUCT
{ omim1} <http://bio2rdf.org/omim#xGeneticDisorder>
{ omim2}
,{omim1} rdfs:label {label1}
,{omim2} rdfs:label {label2}
FROM
{ omim1} <http://bio2rdf.org/omim#xGeneticDisorder>
{ omim2}
,{omim1} rdfs:label {label1}
,{omim2} rdfs:label {label2}
```

Once downloaded, the RDF graph was produced in turtle format and it was possible, with Welkin tool, to draw the network of the 111 interrelated nodes linked by 304 edges (Figure 2). The clustering function of Welkin could, after filtering the network, show only the highly clustered nodes. Obviously the one most connected was omim:168600, the hub record about Parkinson's disease and the PARK1 and PARK8 gene records. It is interesting to see that Alzheimer's disease's hub record is also on the map.

To do a fast literature review of Parkinson's disease, we wanted to identify the most frequently used MeSH terms annotating the Parkinson's papers available on PubMed. To do so, we needed a sample of Parkinson's literature. By submitting this URL to Elmo we asked the crawler to fetch NCBI's PubMed database for the 2,269 papers referenced in the 166 OMIM records. NCBI policy about using efetch to obtain many documents from their server is published at NCBI Entrez Programming Utilities [32]. It restricts the maximum number of queries that can be submitted.

```
http://localhost:8080/bio2rdf/load:pubmed
```

After downloading from NCBI, more than 2000 abstracts were obtained along with their own annotations, weighing 22 Mb/246 Kt. The MeSH annotations used to annotate Parkinson's scientific literature are selected by submitting the following SeRQL query :

```
SELECT *
FROM
{ omim} rdf:type
{<http://bio2rdf.org/omim#GeneticDisorder>}
,{omim} rdfs:label {literal}
,{omim} <http://bio2rdf.org/bibtex#xArticle> {xArticle}
,{xArticle} <http://bio2rdf.org/bio2rdf#xMeSH> {term}
WHERE
literal like "*PARKINSON*"
```

After having exported the results in tabular format, it was possible to further analyze with spreadsheet pivot table tool to filter, count and sort it as shown in Figure 3. Of course "Parkinson Disease" MeSH term is one of the most referenced, but at the top of the list there are also the following interesting hints for a newcomer in the domain : Nerve Tissue Proteins, Brain and Dementia. Such compilation of literature about Parkinson scientific research domain is obtained with a minimum of manipulations and readings.

	A	B
1		
2	Count of omim	
3	term	Total
4	http://bio2rdf.org/mesh:Humans	192
5	http://bio2rdf.org/mesh:Parkinson Disease	150
6	http://bio2rdf.org/mesh:Male	131
7	http://bio2rdf.org/mesh:Female	127
22	http://bio2rdf.org/mesh:Nerve Tissue Proteins	24
23	http://bio2rdf.org/mesh:Parkinsonian Disorders	23
24	http://bio2rdf.org/mesh:Genotype	22
25	http://bio2rdf.org/mesh:Brain	20
26	http://bio2rdf.org/mesh:DNA Mutational Analysis	20
27	http://bio2rdf.org/mesh:Adolescent	19
28	http://bio2rdf.org/mesh:Amino Acid Sequence	19
29	http://bio2rdf.org/mesh:Base Sequence	18
30	http://bio2rdf.org/mesh:Family Health	18
31	http://bio2rdf.org/mesh:Dementia	17
32	http://bio2rdf.org/mesh:Genes, Dominant	17
33	http://bio2rdf.org/mesh:Phenotype	17
34	http://bio2rdf.org/mesh:Polymorphism, Genetic	17

Figure 3: Frequency analysis of MeSH terms found in 2058 PubMed papersabout Parkinson's disease and referenced by OMIM

Our last example shows the power of a Bio2RDF mashup system to aggregate knowledge from very different knowledge spaces. To answer the last question we needed to integrate knowledge from 7 different websites : OMIM, GeneID, UniProt, Ligand and ChEBI about molecules and, finally, Kegg and Reactome about pathways. A first query in our actual knowledge base returned 12 GeneID descriptions related to 9 OMIM records containing the term "parkinson" in their title. These genes were related to 10 pathway definitions from Kegg and 5 from Reactome. By submitting the queries hereafter to Elmo, the program crawled Kegg and Reactome knowledge-bases to obtain their corresponding pathway definitions; Ligand and ChEBI to fetch molecule descriptions; and

finally, GeneID for the involved genes annotations and UniProt for the annotations of proteins associated to the involved genes.

```
http://localhost:8080/bio2rdf/load:keggpathway
http://localhost:8080/bio2rdf/sameas:hsa-geneid
http://localhost:8080/bio2rdf/learn:geneid
http://localhost:8080/bio2rdf/load:cpd
http://localhost:8080/bio2rdf/load:reactome
http://localhost:8080/bio2rdf/load:biopax-xref
http://localhost:8080/bio2rdf/load:chebi
http://localhost:8080/bio2rdf/load:obo-xref
http://localhost:8080/bio2rdf/sameas:keggcompound-cpd
```

Within five minutes of download time from Bio2RDF server, Elmo has instantiated triples in a knowledge base which contained all that is known about pathways from the Kegg KGML representation and from the Reactome BioPAX; more than 2,000 RDF documents from 7 different web sites were loaded into a local Sesame triplestore weighing 29 Mb/546 Kt. The pathway descriptions were enriched with the descriptions of gene, protein and molecule. Since GeneID genes and UniProt proteins were linked, as well as, Ligand's Compound molecules and ChEBI's, it was possible, with an SeRQL query, to analyze how closely the concepts present in both pathway definitions were actually related. The knowledge base contains more than 1,500 GO terms and 14,000 PubMed references. Figure 4 gives a visual representation of the global knowledge base obtained. A node represents a namespace and the number of RDF documents belonging to it. An edge indicates the number of resources linked by URIs between two different namespaces.

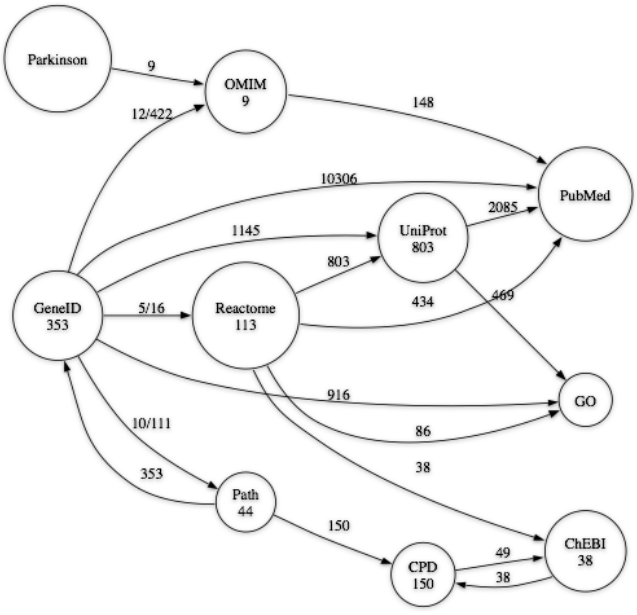

Figure 4: Global knowledge map of Kegg and Reactome pathways containing genes related to Parkinson's disease.

5.DISCUSSION

Bio2RDF is a Work in Progress

Bio2RDF ontology and its rdfizer programs are not definitive. RDF document format will still evolve. We invite interested bioinformaticians to join the bio2rdf.sourceforge.net project. There are many more rdfizers to be written. The bio2rdf.owl ontology is just at an early stage of development, it now needs to be adopted by the community, as was the case for the BioPAX [15] ontology. An ontology belongs to a community who adapts it, uses it and shares it. The Bio2RDF community has still to emerge.

URI Normalization

In this project we created RDF documents from different sources. They all use a simple URI normalization implemented to ease the recombinant effect described in the Material and Method section. In BioPAX [15] and OBO [35] ontologies, there is a triple for the database and another one for the id. This representation does not give a unique URI for each document. To correct this problem, we created two scripts, load:biopax-xref and load:obo-xref, which add to the triple by concatenating DB and ID and creating a normalized URI. Another problem comes from the use of different namespaces to identify the same database source. For example, the namespace for OMIM [14] was identified by "mim" or by "omim". We created a script, called sameas:NSfrom-NSto, to correct this situation. This script adds nodes to link triples belonging to synonym namespaces with the owl:sameAs predicate.

LSID vs URL

There are currently two ways to provide a resolution for the URI of a unique entity: LSID and URL. LSID URI can be resolved and browsed with the FireFox plugin named LSID Browser. URL is readily usable under any browser without any plugin. The current discussion objective in the community is to decide which URI syntax should be used. In the Bio2RDF project it was decided to offer both representations. With the URLrewrite library, we have created a rule to trigger the LSID web services when a query for a LSID is detected. The other rules trigger the usual URL query for a document. This way, Bio2RDF is ready for whatever the community will decide to choose as the standard. The following URI example illustrates the geneid:15275 entity referenced with the URL or LSID form.

```
http://bio2rdf.org/geneid:15275
http://bio2rdf.org/urn:lsid:bio2rdf.org:geneid:15275
```

Scalability of Complexity

In the near future, there will be much more knowledge available to the scientific community coming from different sources and with increasing complexity. How will data be integrated without using a strategy to keep complexity constant in the underlying system? This is the most important property of the RDF framework and, especially, the most useful characteristic of a triplestore. Without a triplestore, RDF documents are just XML. It is inside the triplestore that the inherent recombining characteristic of URIs becomes available once they are normalized. The complexity of the knowledge stored in it can grow without any extra programming to manage it. RDF is a framework that enables a very simple thing: scalability of the knowledge base complexity. The Bio2RDF project proposes to keep complexity in the bioinformatics knowledge space under control by applying this proven web semantic approach.

Use case

With the preceding use case about Parkinson's disease, we have shown the potential of a knowledge framework to build a knowledge-base tailored to a very specific problem. With the warehouse stored into a triplestore, it is possible to query the local knowledge base with SeRQL queries. However, the full promise of the semantic web is to build a truly distributed knowledge-base in the world-wide-web. This will be possible when all

bioinformatics resources are made available in RDF format, and through the use of languages and protocols such as SPARQL (), a standard defined by the W3C to express queries across distributed RDF data sources. The Bio2RDF project, through its combination of warehouse and on-demand tools, addresses the current lack of both RDF data and ontology standards in the bioinformatics community.

6.CONCLUSION

In the Bio2RDF project our main goals were to provide the scientific community with access to normalized RDF documents from many different sources and to offer a method to allow users to add knowledge sources by creating new rdfizers.

We have shown that the semantic Web approach for automatic aggregation of knowledge is very promising. Other research projects have explored data integration with the same approach, but Bio2RDF showed that it is possible to scale up to millions of documents. We succeeded with 46 millions documents coming from twenty different data sources. Since excellent software, dedicated for use with RDF, already exists, creating a friendly user interface to query our networked data was a secondary concern. Despite the ongoing need for user-friendly interfaces to the Bio2RDF service, semantic Web tools working with RDF are blooming and in rapid evolution. Our message to the bioinformatics community is the following: good work can already be done with current semantic Web software and more effort should be directed to providing quality RDF data.

Since we now have access to large amounts of RDF files from biological databases we will study the underlying graph created by linking them together. How much is known? Are there any connections between two entities in the knowledge space of bioinformatics? Can Bio2RDF be used for knowledge discovery?

By giving access to a knowledge space with well organized data in the semantic Web of life sciences, we believe that Bio2RDF is an example of tool that can help to eliminate some of the social hurdles aka. creeps [12] to the adoption of this valuable technology.

7.ACKNOWLEDGMENTS

The Bio2RDF software project was made possible because of the availability of great software from the open source community. Our first thanks go to programmers. It was possible to create the Bio2RDF service because the data providers make huge amounts of curated knowledge publicly available to the biologist community. We also thank them, especially the curators without whom knowledge tagging would not be a reality. Finally, we would like to thank the reviewers for their suggestions.

François Belleau was a recipient of a studentship from Génome Québec and Marc-Alexandre Nolin was a recipient of a studentship from the Canadian Institutes of Health Research. This work has been financed, in part, by the Atlas of Genomic Profiles of Steroid Action, a Genome Canada project.

1.REFERENCES

[1] Aduna, Sesame, http://www.openrdf.org

[2] Aduna, Elmo, http://www.openrdf.org

[3] Ashburner, M. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., http://dx.doi.org/10.1038/75556 Nat Genet, 2000, 25, 25-29

[4] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. & Bourne, P. E., The Protein Data Bank., Nucleic Acids Res, 2000, 28, 235-242

[5] Berners-Lee, T., Building on what exists, http://www.w3.org/2006/Talks/0404-mit-tbl 2006

[6] Brickley, D. & Miller, L., FOAF Vocabulary Specification, http://xmlns.com/foaf/0.1/

[7] Cheung, K. et al., YeastHub: a semantic web use case for integrating data in the life sciences domain., http://dx.doi.org/10.1093/bioinformatics/bti1026 Bioinformatics, 2005, 21 Suppl 1, i85-i96

[8] Davidson, S. B.; Overton, C. & Buneman, P., Challenges in integrating biological data sources., J Comput Biol, 1995, 2, 557-572

[9] Eppig, J. T. et al., The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology., Nucleic Acids Res, 2005, 33, D471-D475

[10] Fox, J. A.; McMillan, S. Ouellette, B. F. F., A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory., http://dx.doi.org/10.1093/nar/gkl379 Nucleic Acids Res, 2006, 34, W3-W5

[11] Fujibuchi, W. et al., DBGET/LinkDB: an integrated database retrieval system., Pac Symp Biocomput, 1998, 683-694

[12] Good, B. M. & Wilkinson, M. D., The Life Sciences Semantic Web is full of creeps!, http://dx.doi.org/10.1093/bib/bbl025 Brief Bioinform, 2006, 7, 275-286

[13] Gruber, T., Toward Principles for the Design of Ontologies Used for Knowledge Sharing, International Journal of Human and Computer Studies, 1995, 43, 907-928

[14] Hamosh, A. et al., Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders., http://dx.doi.org/10.1093/nar/gki033 Nucleic Acids Res, 2005, 33, D514-D517

[15] Hogue, C.; Karp, P. & Sander, C., BioPAX : Biological Pathway Exchange, http://www.biopax.org

[16] Hubbard, T. J. P. et al., Ensembl 2007., http://dx.doi.org/10.1093/nar/gkl996 Nucleic Acids Res, 2007, 35, D610-D617

[17] Hulo, N. et al., The PROSITE database., http://dx.doi.org/10.1093/nar/gkj063 Nucleic Acids Res, 2006, 34, D227-D230

[18] Huynh, D.; Mazzocchi, S. & Karger, D., Piggy Bank: Experience the Semantic Web Inside Your Web Browser, 2005

[19] IBM, LSID (Life Sciences Identifier) Resolution Project, http://www.omg.org/cgi-bin/doc?lifesci/2003-12-02

[20] Jain, E., Uniprot RDF, http://dev.isb-sib.ch/projects/uniprot-rdf/

[21] Joshi-Tope, G. et al., Reactome: a knowledgebase of biological pathways., http://dx.doi.org/10.1093/nar/gki072 Nucleic Acids Res, 2005, 33, D428-D432

[22] Kanehisa, M. et al., From genomics to chemical genomics: new developments in KEGG.,

http://dx.doi.org/10.1093/nar/gkj102 Nucleic Acids Res, 2006, 34, D354-D357

[23] Knouf, N., bibTeX Definition in Web Ontology Language (OWL) Version 0.1, Working Draft, http://zeitkunst.org/bibtex/0.1/ 2004

[24] Köhler, J.; Philippi, S. & Lange, M., SEMEDA: ontology based semantic integration of biological databases., Bioinformatics, 2003, 19, 2420-2427

[25] Maglott, D.; Ostell, J.; Pruitt, K. D. & Tatusova, T., Entrez Gene: gene-centered information at NCBI., http://dx.doi.org/10.1093/nar/gkl993 Nucleic Acids Res, 2007, 35, D26-D31

[26] Mazzocchi, S. & Ciccarese, P., SIMILE | Welkin, http://simile.mit.edu/welkin/

[27] Schuler, G. D.; Epstein, J. A.; Ohkawa, H. & Kans, J. A., Entrez: molecular biology database and retrieval system., Methods Enzymol, 1996, 266, 141-162

[28] Baker C. J. O., Shaban-Nejad A., Xu S., Haarslev V. and Butler G., Semantic Web Infrastructure for Fungal Enzyme Biotechnologists, Journal of Web Semantics, (4) 3, 2006.

[29] Stein, L. D., Integrating biological databases., http://dx.doi.org/10.1038/nrg1065 Nat Rev Genet, 2003, 4, 337-345

[30] Stevens, R.; Baker, P.; Bechhofer, S.; Ng, G.; Jacoby, A.; Paton, N. W.; Goble, C. A. & Brass, A., TAMBIS: transparent access to multiple bioinformatics information sources., Bioinformatics, 2000, 16, 184-185

[31] Mashup (web application hybrid), http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)

[32] NCBI User system requirements, http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html#UserSystemRequirements

[33] The Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/

[34] The Dublin Core Metadata Initiative, http://dublincore.org/

[35] Open Biomedical Ontologies, http://obo.sourceforge.net/

[36] BioRDF subgroup of the HCLS community, http://www.w3.org/2001/sw/hcls/

[37] http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary

[38] http://dev.isb-sib.ch/projects/uniprot-rdf/migration.html

[39] http://enki.isb-sib.ch/projects/uniprot-rdf/data/core.owl

[40] http://www.cs.concordia.ca/FungalWeb/OntoIQ.html

[41] http://www.racer-systems.com

[42] http://java.sun.com/products/jsp/jstl/

[43] http://tuckey.org/urlrewrite/

[44] http://bio2rdf.org/bio2rdf-2007-02.owl

[45] http://simile.mit.edu/RDFizers/

[46] http://www.berkeleybop.org/ontologies/