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.

2. Materials and methods

Two main ideas have oriented our software development: the conversion of existing databases into RDF format (a process called "rdfizing") and the use of existing semantic web software to merge, query and visualize the data. These software components are: the Sesame open source triplestore, the Protégé ontology editor, the Piggy Bank [32] semantic web browser plug-in for FireFox and the Welkin [33] RDF graph visualizer, both developed by the MIT, and finally the experimental LSID browser [34].

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

2.1. Ontology design

An ontology can be defined as an explicit specification of a conceptualization, a conceptualization being an abstract and simplified view of the world that needs to be represented for some purpose [35,13]. For a given knowledge base or knowledge system, it means that a conceptual language should be used to define the objects and the relations to be represented. OWL is the conceptual language chosen by the semantic web community for ontology-based knowledge representation. To design the ontology of Bio2RDF, we used the Protégé open source framework and its OWL editor Protégé-OWL.

Since the main goal of Bio2RDF was to convert into RDF format the documents available on the web (Entrez Gene description of Hk1 from NCBI website for instance), the first step was 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 the URI of the resource, usually defined in another namespace like GI, GO or PubMed. Using this approach we produced an OWL description from each selected HTML document. This step was repeated for each namespace recognized by the current version of Bio2RDF: GO, OMIM, PDB, etc. For BioPAX and UniProt, this step was unnecessary because their OWL schema was already available. Finally, the global bio2rdf-2007-02.owl [36] ontology description was built by merging the ontology file of each namespace.

After the Bio2RDF ontology was created, the second step consisted in 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 syntax. The creation of rdfizer programs, performed for more than twenty different namespaces, was the main task of the Bio2RDF project.

The design of the Bio2RDF's ontology was inspired by already existing ontologies. For instance, rdf:type and rdfs:label were systematically used in each document. The label predicate always contains the name of the resource followed by a short form of its URI enclosed with "[ ]". For example, rdfs:label of geneid:15275 is "hexokinase 1 (Hk1) [geneid:15275]". Some common predicates from the Dublin Core project [37] were used, in particular dc:title, dc:identifier, dc:created and dc:modified. We also used the FOAF [38] namespace to describe people and the bibTeX [39] namespace for literature references. We had to create our own predicates in the bio2rdf namespace, the most frequently used ones being bio2rdf:url, bio2rdf:urlImage, bio2rdf:xRef, bio2rdf:name and bio2rdf:synonym. The definition of the semantics of these predicates can be found in the Bio2RDF ontology file [36].

2.2. Rdfizer programs

In an ideal world, all the data would be available in RDF format with complete normalization of URIs, and all documents on Internet would consequently connect together automatically. But this is not the case in the real world. At this time, what exists is an HTML version of the data accessible through web pages. The Bio2RDF project provides RDF formatted documents from several data sources in a normalized way. A JSP toolbox has been created to generate RDF files from locally stored databases or directly from HTML documents accessed via http requests. JSP tools were used to create rdfizers, which are programs transforming existing data into an RDF representation [40]. Several different sources of data can be rdfized: relational databases, text files, XML documents, and HTML pages. For each type of knowledge source, a JSP program converts the data from the original source into the RDF format. These programs use XPath, regular expressions or SQL queries to extract knowledge from the original data. For example:

XML to RDF conversion with an OMIM record from NCBI: The program ncbi-omim2rdf.jsp converts the XML representation of OMIM records provided by the NCBI efetch service [41] into RDF. First, this program receives in parameter the OMIM id for the disease under consideration. In the beginning of the program, it fetches the information from the NCBI website and places the XML document in memory where the translation work can be done. Second, the information for the document is extracted according to the ontology previously created. In JSP, the JSTL XML library can be used to navigate across the document retrieved using efetch.

SQL to RDF conversion with an Ensembl record: Ensembl [42] provides online access to its MySQL relational databases. The program ensembl-g2rdf.jsp is used to do the conversion. The JSTL SQL library offers functionalities to work with databases. The first action is to establish a connection to the database using the parameters that the providers (Ensembl in this example) supply on their website. Once the connection is established, queries are created to fetch all the data required to create the corresponding RDF document.

Text file to RDF conversion with a Prosite record: Prosite [43] website returns a text format description of a protein family domain. The rdfizer prosite2rdf.jsp retrieves this text document, then uses regular expressions to parse its content and to generate an RDF version out of it.

The format of the RDF document produced by the Bio2RDF rdfizer is not considered definitive. For this reason, the source code of all our rdfizer programs has been made publicly available for customization.

Although rdfizers on the Bio2RDF.org server or client can be queried directly, they also have a REST-like [44] interface granted by the use of an urlrewrite filter. This allows changing the underlying programs without modifying the query methods, thus providing stable URIs that can deal with changes in URIs by upstream data providers. Such stable URIs are critical for properly linked data and this subject is further elaborated in the following section.

Bio2RDF is a three-step approach elaborated and tested for a mashup of bioinformatics data. The first step is to build a list of namespaces for different data providers. This enables the construction of normalized URIs. The second step is to analyze a data source to represent it in the RDF model. The third step is to build an rdfizer that converts the information from the data source into its RDF representation. The resulting RDF documents can then be put into a triplestore in order to connect together. Further analysis can be made on the triplestore with SeRQL or "REST like" queries.

2.3. URI normalization

The availability of RDF documents is not by itself sufficient to obtain a mashup. External references, expressed as URIs, need to be normalized to allow proper connection of triples. For example, a PubMed reference with an 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 a functional connection between triples. Even in existing well-formed RDF documents there is a problem with URIs. For example the GO term of Hexokinase (GO:0004396) is referenced by different URIs used by existing RDF data providers: UniProt, OBO and BioPathways Consortium.

http://www.geneontology.org/go#GO:0004396
http://purl.uniprot.org/go/0004396
urn:lsid:geneontology.org.lsid.biopathways.org:go:0004396

Those are all supposed to represent the same concept: the definition of Hexokinase molecular function according to Gene Ontology. If we were to load these documents and use them in the same triplestore, no links would be created around the Hexokinase concept because the URIs are all different even if they correspond to the same concept. By adopting the same URI pattern for all URIs regardless the provider, the Bio2RDF system guarantees that the connections are built automatically around the same concept. The Bio2RDF URI synonym for the preceding published URI is:

http://bio2rdf.org/go:0004396

The Bio2RDF's global strategy to ensure that the RDF graph refers to unique concepts is accomplished by applying the Bio2RDF URI syntax wherever possible. When an URI has already been assigned to a graph by the data provider, we keep track of it by adding an owl:sameAs predicate linking to this official URI.

Many proposals revolve around this subject, each and every one having its pros and cons. The LSID proposal [14] is an identification scheme using SOAP for content negotiation. The scheme makes it possible to keep a stable URI even if the provider disappears because it is domain independent, but at the cost of not being a routable identifier. Another scheme is content negotiation with 303 redirections, which would give routable URIs, but this behavior is not the default one for a web server and the client has to be built to ask for XML/RDF content type or else they will receive the HTML page instead of the RDF one. The Bio2RDF project has established a simple set of rules that data providers can apply to create URIs for their information:

1. *Use a REST like interface*. REpresentational State Transfer (REST) enables us to produce a clear and stable URI for every document. A default action can also be used, but it must be explained on the data provider's website. Either way, the data provider should create a web page explaining her REST interface. Also, a REST like interface does not need content negotiation, web application or server redirection.
2. *Lowercase all the URI up to the colon*. The URI case sensitivity poses a problem because each different case results in a theoretically different URI. The two most successful kinds of URI are domain names and email identifiers, which are case insensitive. This allows addresses differing solely by case (such as uniprot.org and UniProt.org) to refer to the same site. We suggest converting into lower case the URI part up to the colon, rendering it effectively case insensitive.
3. *All URIs should return an RDF document*. If an URI for a document returns a web page instead of an RDF document, it is not easy to connect its information directly with other linked data. Transformation will need to be done before the RDF graph is obtained. This very important rule is equivalent to rule #2 of the Linked data design rules from the W3C [45]: "*Use HTTP URIs so that people can look up those names*." According to this rule, using a HTTP *dereferencable* URI to identify a resource is natural and very simple to use because the URI returns (dereference) an RDF graph. Usage of LSID does not respect this specific rule because there is no reference to any protocol in its URI.

Rules to convert URIs have been adopted for this reason. The major one is that a URI is uniquely attributed to a document which describes an object like an ontology term, a gene description or a protein annotation. 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 the 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, the triples referring to this document connect to it.

2.4. Bio2RDF architecture

Fig. 1 shows a schematic description of the Bio2RDF architecture. All external data sources, in different formats (XML, Text, ASN.1, KGML and RDF), are listed on the left part. These sources are processed in two different ways. Websites, from which the entire database was downloaded (MGI [46], HGNC [47], Kegg [12], Entrez Gene [27], OMIM [26], GO [28], OBO [48], PDB [49] and ChEBI [50]) to the Bio2RDF.org server, form the top group. 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 is stored in a MySQL database. Only requested documents from websites in the bottom group (UniProt, Reactome, Prosite, PubMed, GenBank, PubChem, etc.) are rdfized directly from the original source. In 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 from Entrez Gene, OMIM, OBO and Kegg, is cached for availability and speed purpose: availability because few data providers have an RDF version of their documents, and speed because some data providers have restrictions for the access of documents.

The new http://beta.uniprot.org server now offers a RDF graph accessible by dereferencing HTTP URI of the form:

http://purl.uniprot.org/DATABASE/ID.rdf

Here *DATABASE* can be any of UniProt Consortium main databases: uniprot, uriref, uniparc, etc. With millions of RDF documents available from this single source at high speed, the only transformation done by the Bio2RDF service consists in replacing the syntax of URI. Ultimately, the data should always come live from the data providers as done now with this new UniProt RDF service. Then the Bio2RDF server would only act as a proxy server forwarding the RDF query from the client requesting a graph to the real data provider.

The myBio2RDF application contains two servlets running under a Tomcat server: Elmo and Sesame. Elmo is a RDF crawler which was originally created to follow rdfs:seeAlso predicate in-

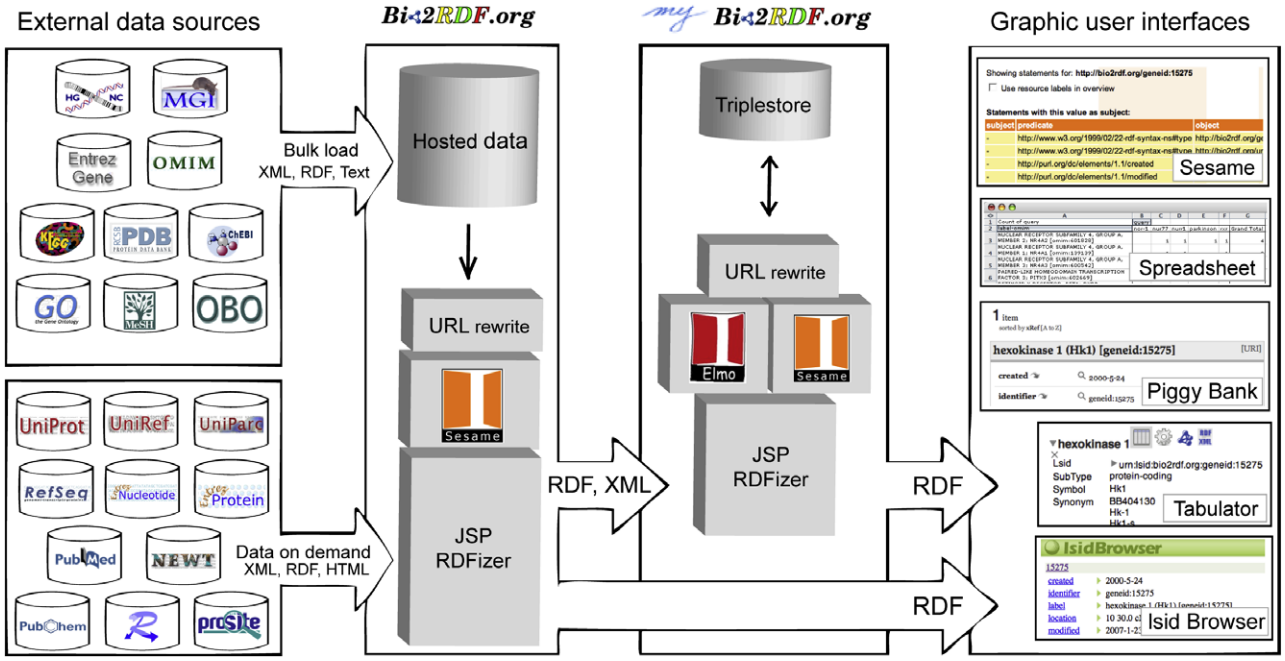

Fig. 1. Bio2RDF knowledge system framework architecture.

cluded 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. Three specific services were added to allow Elmo to crawl specific knowledge:

(1) To obtain a list of URIs corresponding to the results of a text search using the search engine of the corresponding website. This tool is very useful because it leverages the existing text search capability available from the official data provider.
http://localhost:8080/search:TEXT@database
where database = [omim|geneid|pubmed|mesh|kegg|uniprot]
(2) To request all URIs in the triplestore which belongs to the specified namespace.
http://localhost:8080/load:NAMESPACE
(3) To create a synonym node to link two URIs which have the same id but different synonymous namespaces. For example, to link omim:602080 and mim:602080 URI together because the *omim* namespace is equivalent to *mim*.
http://localhost:8080/sameas:NAMESPACE1-NAMESPACE2

The URLRewrite library matches the URL syntax with the appropriate RDFizer program. This software component controls the information workflow by interpreting rules defined by regular expressions. Examples of rules stored in the URLRewrite configuration file follow. 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>
```

2.5. Resulting REST-like URL

http://bio2rdf.org/pubmed:12728276

The next rule forwards to the 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, just like DNS servers do.

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

When a query is made to the Bio2RDF service for an unknown URI, for example http://bio2rdf.org/biocyc:MONOMER-9282 for which there is no rdfizer program and no URLRewrite rule, the server responds with a graph of type bio2rdf:Unknown.

The flexible Bio2RDF approach allows the replacement of one rdfizer by another simply by modifying the corresponding URL rewrite rule. It is also possible to add a new rdfizer program working with private data locally stored in a relational database. Once a new extension is added, new knowledge sources can then be merged with Bio2RDF. This is the way that the myBio2RDF application learns how to explore a new knowledge space.

3. Results

The Bio2RDF project is still in development, but its RDF document service has been made available to the scientific community. More than twenty different public bioinformatics data sources are now available in a 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 from where documents were converted by the rdfizer programs.

These databases are NCBI's Entrez Gene and OMIM, Kegg's pathway and Ligand, MGI mouse's annotations and HGNC human's annotations, OBO open source ontology, PDB the Protein Data Bank, and finally ChEBI the chemical entities database from EBI. The MeSH RDF version of the medical vocabulary comes from previous work done by van Assem [51]. By locally storing these major databases, hundreds of RDF documents, related to a specific topic, can be extracted in minutes rather than hours. It also helps to respect NCBI usage restrictions [52]. Some other knowledge bases are also available in RDF format from the Bio2RDF server, although they are not hosted on it: the UniProt protein knowledge base and its taxonomy that were recently made available in RDF format [53], PubMed, GenBank and PubChem from NCBI accessible with the efetch utility, Reactome and Prosite. As an example, pathway definitions are offered in BioPAX RDF format from the Reactome 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 locally stored in our database. This knowledge space corresponds approximately to 163 million of well-formed RDF documents using normalized URIs and respecting the Bio2RDF ontology.

The databases downloaded had different formats. The UniProt knowledge base was available directly in RDF format. NCBI offered all its main databases in ASN.1 format from its FTP site where the Entrez Gene database was downloaded in ASN.1 format before its conversion into XML format. Recent work by Sahoo [54] has been done at NIH to convert Entrez Gene to RDF but this resource is not publicly available so it cannot be used yet. The OMIM database was available only in tabulated text files so the efetch utility was used to extract each OMIM's record individually in XML format. Gene Ontology was available from three different sources (GO, OBO and UniProt) in three different RDF schemas. The GO's FTP server was chosen because it was the authoritative website. The PDB server releases all its records over an FTP server. Kegg's pathways were downloaded from their FTP server in KGML (http://www.genome.jp/kegg/docs/xml/), an XML proprietary format. The LIGAND database, with documents about compound, reaction and enzyme, could be downloaded only in text format. A Perl program was written to rdfize them. The MGI mouse genome's annotations, originally in tabulated text files, were transformed into RDF the same way. Finally, the OBO's ontologies were downloaded in a RDF version.

With the Bio2RDF server or the myBio2RDF application, it is possible to browse millions of RDF documents using the Sesame explorer to view HTML page, Piggy Bank [32] or the experimental FireFox extensions: LSID browser [34] or Tabulator [55]. The next section explains how an agent can automate this process and the role of the Elmo RDF crawler.

4. Parkinson use case

The potential of the Bio2RDF approach to build specialized mashups is illustrated by applying it to the construction of a knowledge base about Parkinson's disease (PD). This disease was chosen because it was already analyzed by the BioRDF subgroup of the HCLS community [60], and also in reason of the availability of a Parkinson's disease specialist at the CHUL Research Center, Claude Rouillard. The following paragraph explains part of his research.

Parkinson's disease is a slow progressive neurodegenerative disorder. Most cases of PD are sporadic, but rare familial forms of the disease do occur. However, the mechanisms underlying the selective death of nigral dopamine (DA) neurons are still unknown. Nuclear receptors constitute a conserved family of ligand activated transcription factors regulating gene expression. We and others have provided several lines of evidence suggesting an important involvement of a subgroup of nuclear receptors specifically associated with DA neurotransmission in the developing and mature CNS. This subgroup includes the Retinoid X Receptor (RXR) and orphan members of the thyroid/steroid nuclear receptor family named the Nur family, which includes Nurr1, Nur77, and Nor-1. Nurs are classified as early response genes, and are induced by diverse signals, including growth factors, cytokines, peptide hormones, neurotransmitters, and stress. Their ability to sense and rapidly respond to changes in the environment seems to be a hall-

Table 1
Number of RDF documents from public databases available with Bio2RDF

Data source	Short URI example	Number of RDF documents	Format of source data	Hosted version
genenames.org	hgnc:4922	27,634	Tabulated	December 2007
informatics.jax.org	mgi:96103	70,172	Tabulated	June 2007, MGI 3.54 release
ncbi.nlm.nih.gov	omim:146200	18,284	XML	December 2007
ncbi.nlm.nih.gov	geneid:3098	3,315,893	XML	December 2007
genome.ad.jp	path:mmu00010	68,307	KGML	December 2007, Release 44.0+/12-19
genome.ad.jp	cpd:C00011	15,006	Text	December 2007, Release 44.0+/12-19
genome.ad.jp	dr:D00001	6755	Text	December 2007, Release 44.0+/12-19
genome.ad.jp	ec:2.7.1.1	4,958	Text	December 2007, Release 44.0+/12-19
genome.ad.jp	gl:G00001	10,972	Text	December 2007, Release 44.0+/12-19
genome.ad.jp	rn:R00014	7422	Text	December 2007, Release 44.0+/12-19
ebi.ac.uk	chebi:16526	13,360	Tabulated	December 2007, Release 39.0
rcsb.org	pdb:1HKC	48,091	XML	December 2007
geneontology.org	go:0004396	24,634	OBO/RDF	December 2007
nlm.nih.gov	mesh:D006593	23,512	RDF	Febuary 2007
obofoundry.org	*obo's 54 namespaces*	108,955	OBO/RDF	December 2007
beta.uniprot.org	uniparc:UPI00005AC213	30,261,843	RDF	
beta.uniprot.org	uniprot:P19367	4,177,176	RDF	
beta.uniprot.org	uniref:UniRef50_P19367	7,990,452	RDF	
beta.uniprot.org	taxon:9606	441,422	RDF	
ncbi.nlm.nih.gov	genbank:NP_277035	61,132,599	XML	
ncbi.nlm.nih.gov	pubmed:3207429	17,000,000	XML	
ncbi.nlm.nih.gov	pubchem:3313	38,000,000	XML	
reactome.org	reactome:70326	8,332	BioPAX/RDF	
expasy.org	prosite:PS00378	2,819	HTML	
	Total	162,778,598		

mark of this subgroup. Numerous results suggest that impaired Nurr1 function may be associated with an increased vulnerability of dopamine neurons to degeneration in Parkinson's disease whereas both Nur77 and Nor-1 are important signals for apoptosis pathways outside the brain. Interestingly, Nur77 functions as a survival factor in the nucleus whereas it is a potent killer when migrating to the mitochondria.

The mashup created with Bio2RDF about PD will help answer these questions:

1. Which GO terms describe our four genes of interest (Rxr, Nurr1, Nur77, and Nor-1)?
2. Which articles mentioning our four genes of interest are related to apoptosis AND cytoplasm and also mention genes having GO annotations about apoptosis OR cytoplasm?

First, a knowledge base is built to answer these questions. This knowledge base is loaded with the relevant documents from different sources to answer a specific question.

```
01   SELECT DISTINCT
02     searchLabel, geneLabel, goLabel
03   FROM
04   {search} rdf:type {<http://bio2rdf.org/bio2rdf#Search>};
05            <http://bio2rdf.org/bio2rdf#query> {searchLabel};
06            rdfs:seeAlso {gene},
07   {gene}   rdfs:label {geneLabel};
08            <http://bio2rdf.org/bio2rdf#xGO> {go},
09   {go}     rdfs:label {goLabel}
```

The data mashup building procedure can be reproduced using the myBio2RDF application to build the needed knowledge base. Initially, the related RDF documents are added to the triplestore to our four genes of interest. This is done by submitting the following URIs to the Elmo crawler application:

http://localhost:8080/bio2rdf/search:nur77@geneid
http://localhost:8080/bio2rdf/search:nurr1@geneid
http://localhost:8080/bio2rdf/search:nor-1@geneid
http://localhost:8080/bio2rdf/search:rxr@geneid

This shows the search service provided by Bio2RDF which invokes the NCBI's own Entrez search. The search for Nur77 returns 38 genes, Nurr1 returns 28 genes, Nor-1 returns 17 genes and RXR returns 78 genes. With the next step, we add the PubMed and GO annotations they are referring to are added into the triplestore by submitting the two following URIs:

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

With these documents in the triplestore, we can now try to answer the questions. First, we want to characterize our genes of interest: RXR, Nurr1, Nur77 and Nor-1. Each gene's description from Entrez refers to several Gene Ontology identifiers. The load:go URI is used to fetch the identifier's description. With a SeRQL query we gather all these GO terms about our genes of interest and the result is transferred to a spreadsheet to create a cross table out of it. The query returns 385 distinct GO terms and a total amount of 1295 annotations for all of the four genes.

Line 2 specifies the three columns of the result. The first column, searchLabel, contains the name of the terms that were looked for in NCBI's Entrez. The second column contains the name of the gene related to one of the four genes we searched for. The third column contains the GO term name. The result of this query, based on documents from two different sources (GO and Entrez Gene), is shown in Table 2.

In Table 2, two GO terms, cytoplasm and apoptosis, are highlighted because they are involved in the second question. We choose these terms because Nur77 mediated apoptosis outside the brain involves its translocation from the nucleus to the cyto-

Table 2
GO terms frequency for four genes of interest related to Parkinson's disease

GO terms	Genes of interest				
	Nor-1	Nur77	Nurr1	RXR	Total
Nucleus [go:0005634]	7	14	12	28	61
Regulation of transcription, DNA-dependent [go:0006355]	7	10	10	21	48
Protein binding [go:0005515]	3	11	9	21	44
Metal ion binding [go:0046872]	4	9	8	21	42
Transcription [go:0006350]	5	10	7	20	42
Transcription factor activity [go:0003700]	6	7	8	21	42
Zinc ion binding [go:0008270]	4	9	6	20	39
Sequence-specific DNA binding [go:0043565]	6	6	8	18	38
Steroid hormone receptor activity [go:0003707]	4	5	5	17	31
Signal transduction [go:0007165]	3	11	6	10	30
Positive regulation of transcription from RNA polymerase II Promoter [go:0045944]	3	6	4	10	23
Cytoplasm [go:0005737]		**7**	**4**	**7**	**18**
DNA binding [go:0003677]	1	3	5	9	18
...					
Anti-apoptosis [go:0006916]		1	2	1	4
Apoptosis [go:0006915]		**3**		**1**	**4**
Biological_process [go:0008150]	1	1	1	1	4
...					

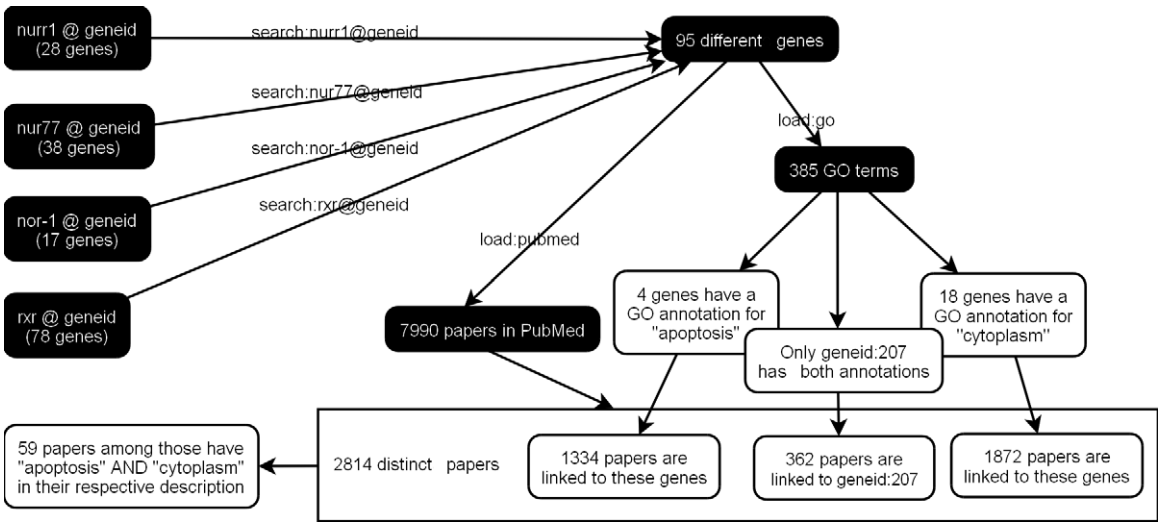

Fig. 2. Number of documents and subsequent restrictions on the knowledge base.

plasm. Once again, we can answer this more complicated question with a SeRQL query over the same knowledge base.

```
01   SELECT DISTINCT
02     geneSearch, geneLabel, goLabel, articleLabel
03   FROM
04   {search}  rdf:type {<http://bio2rdf.org/bio2rdf#Search>};
05             <http://bio2rdf.org/bio2rdf#query> {geneSearch};
06             rdfs:seeAlso {gene},
07   {gene}    rdfs:label {geneLabel};
08             <http://bio2rdf.org/bio2rdf#xArticle> {article};
09             <http://bio2rdf.org/bio2rdf#xGO> {go},
10   {go}      rdfs:label {goLabel},
11   {article} rdfs:label {articleLabel};
12             p {literal}
13   WHERE
14   (
15     go = <http://bio2rdf.org/go:0006915>
16   OR
17     go = <http://bio2rdf.org/go:0005737>
18   )
19   AND
20   (
21     literal like "*apoptosis*" ignore case
22   AND
23     literal like "*cytoplasm*" ignore case
24   )
```

Genes are the starting point of the knowledge base graph. Lines 8 and 9 show that genes are linked to GO terms and PubMed articles. Line 6 selects genes that come from Entrez Gene. We want to restrict the search for articles to genes having either "apoptosis" or "cytoplasm" in their GO annotations. This restriction is specified on lines 15 and 17 where http://bio2rdf.org/go:0006915 is the URI for "apoptosis" and http://bio2rdf.org/go:0005737 is the URI for "cytoplasm". Finally, with line 12, we do a full text search over all text (title, abstract and MeSH annotations) of PubMed articles for literals specified on lines 21 and 23. The Fig. 2 illustrates the building and the querying of the mashup. Each black node corresponds to a building step of the knowledge base when documents were added to the triplestore. White nodes correspond to restrictions in the second SeRQL query.

Finally, we have transferred the query result into a spreadsheet in order to create a last cross table. Table 3 depicts the relationship between the genes of interest and various apoptosis-related factors. Some of the related links are relevant to the mechanisms by which the genes of interest might be involved in dopamine neuron degeneration.

5. Discussion

5.1. URI normalization

In this project. we have created RDF documents from many different sources, implementing a simple URI normalization scheme to solve the recombinant effect described in the Material and Method section. The good usage of URIs is a central issue in RDF bioinformatics databases. Providers such as UniProt have now replaced LSIDs (used in the development phase of the beta. uniprot.org project) by HTTP URIs. Bio2RDF has adopted the same

Table 3
Article frequency of related genes, containing apoptosis and cytoplasm in their abstract, and related to our genes of interest

Genes of interest	Related genes	GO terms		
		Apoptosis [go:0006915]	Cytoplasm [go:0005737]	Total result
Nur77	BCL2-like 11 (apoptosis facilitator) (Bcl2l11) [geneid:12125]	2		2
	Cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4) (CDKN2D) [geneid:1032]		1	1
	Histone deacetylase 7A (HDAC7A) [geneid:51564]		2	2
	Nur77 downstream gene 1 () [geneid:368204]		1	1
	Tumor necrosis factor (TNF superfamily, member 2) (TNF) [geneid:7124]	4		4
	v-akt murine thymoma viral oncogene homolog 1 (AKT1) [geneid:207]	15	15	30
Nurr1	secreted phosphoprotein 1 (Spp1) [geneid:20750]		1	1
RXR	B-cell leukemia/lymphoma 2 related protein A1a (Bcl2a1a) [geneid:12044]	1		1
	Caspase 8, apoptosis-related cysteine peptidase (CASP8) [geneid:841]		22	22
	Nuclear receptor coactivator 2 (NCOA2) [geneid:10499]		1	1
	v-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3, p65 (avian) (RELA) [geneid:5970]		10	10
	Total result	22	53	75

approach and we hope to see other database providers make their data available in RDF with a similar service based on dereferencable URIs by HTTP queries.

5.2. Compatibility with ongoing Semantic web projects

By designing Bio2RDF according to the linked data rules [49], we have created a knowledge space directly usable by a true semantic web browser such as Tabulator, in order to browse the knowledge space of bioinformatics and define the queries dynamically based on the path traveled. Bio2RDF is in use in the demo section of the Tabulator [56]. This demo shows the usefulness of linked data with normalized URIs from different databases.

The Bio2RDF RDF graph can also be browsed with a LSID browser such as [34] through a SOAP web service [57] to request the RDF graphs by LSID. When using this service, the Bio2RDF URIs are replaced by LSIDs with bio2rdf.org for domain name, so http://bio2rdf.org/geneid:15275 becomes urn:lsid:bio2rdf.org:geneid:15275. Like other LSID resolvers do, this URL returns the corresponding graph with LSID in place of URI:

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

Facet browsing is also an important aspect of semantic web application interfaces. Because Bio2RDF returns an RDF graph that can be loaded into the Piggy Bank Semantic web facet browser [32], once a number of graphs of interest have been loaded into its local triplestore, it is possible to do facet browsing in this knowledge space.

Fig. 1 illustrates the use of those different browsing tools that may be employed to browse the semantic web knowledge space available through the Bio2RDF service.

5.3. Extendability

The Bio2RDF architecture was designed with extendability in mind. In addition to the Bio2RDF web services, the myBio2RDF application enables users to integrate local and private data and link them to the Bio2RDF knowledge space. New database sources can easily be added to the system in a few simple steps:

1. Design the RDF document representing the data, using a tool such as Protégé;
2. Write the corresponding rdfizer program to convert the data into a well formed RDF/XML document;
3. Install the new rdfizer program under the Bio2RDF servlet of the myBio2RDF installation;
4. Add a rewrite rule to the urlrewrite.xml configuration file to associate the new rdfizer program to the URI associated to the namespace;
5. Restart the myBio2RDF servlet.

Once a new rdfizer program for a public database has been written, it could be submitted to the Bio2RDF project team for addition to the public Bio2RDF service.

5.4. Scalability of complexity

In the near future, more knowledge will be available to the scientific community, from more 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 contribution of the RDF framework, and the most useful characteristic of a triplestore. Without a triplestore, RDF documents are just XML records. It is inside the triplestore that the inherent recombining characteristic of URIs becomes available if they are normalized. The complexity of the knowledge stored in the triplestore 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 semantic web approach.

5.5. Use case

With the former use case about Parkinson's disease, we have shown the potential of the Bio2RDF knowledge framework to build a very specific knowledge base, the mashup. It was then queried with SeRQL to answer some very specialized questions. The procedure employed in the use case is versatile because, by modifying the SeRQL query, we can search for all kinds of relations between genes of interest and GO terms. It is also very efficient because of the speed and simplicity by which we can gather documents from many different linked data sources providing RDF documents.

5.6. Bio2RDF is a work in progress

The Bio2RDF's ontology and its rdfizer programs are not definitive. The 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 and augmented by the community, as it was the case for the BioPAX ontology. An ontology belongs to a community who adapts it, uses it and shares it. With the warehouse stored into a triplestore, it is possible to query the local knowledge base with SeRQL queries. However, the semantic web is meant to be distributed. With more RDF resources available on the web and by using the SPARQL [58] language and protocol, a standard defined by the W3C, the data warehousing concept could become obsolete in the future. This is one perspective of the semantic web.

6. Conclusion

In the Bio2RDF project, our main goal was to create a framework that could be used to create an on-demand knowledge base to form a mashup of data in the bioinformatics domain. This framework provides an access to normalized RDF documents from many different sources, and offers a method for users to add knowledge sources by creating new rdfizers and also a way to keep privacy of private data by using its built-in routing capability.

We have shown that the semantic web approach for automatic knowledge aggregation is promising. Other research projects have explored data integration with similar approaches but Bio2RDF showed that it is possible to scale up to millions of documents (in our example, 163 million documents from more 20 different data sources). With the availability of software dealing with RDF documents, the elaboration of a friendly user interface to query our networked data was a secondary concern, at least as a first step. Despite the ongoing need for friendly user interfaces to the Bio2RDF service, semantic web tools working with RDF are 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 improve the quality of 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 and apply Bio2RDF to 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 eliminate some of the social hurdles (aka.creeps [59]) to the adoption of this valuable technology.

The myBio2RDF application, which is a modified version of the Sesame triplestore with rdfizers, can be downloaded at http://sourceforge.net/projects/bio2rdf/.

Acknowledgments

The Bio2RDF software project was made possible because of the availability of software from the open source community. Our first thanks go to programmers of this community. It was possible to create the Bio2RDF service because huge amounts of curated knowledge are made publicly available to the biologist community by the data providers. We also thank them, especially the curators without whom knowledge tagging would not be a reality. We also thank Claude Rouillard for his help in the production of the example with Parkinson's disease. Finally, we would like to thank the reviewers for their valuable 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 project funded by Genome Canada and Génome Québec.

This paper is an extension of our workshop paper'Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge System' published in WWW2007/HCLS-DI (http://www2007.org/workshop-W2.php).

References

[1] Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. Methods Enzymol 1996;266:141–62.
[2] Fujibuchi W, Goto S, Migimatsu H, Uchiyama I, Ogiwara A, Akiyama Y, et al. DBGET/LinkDB: an integrated database retrieval system. Pac Symp Biocomput 1998:683–94.
[3] Fox JA, McMillan S, Ouellette BFF. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. Nucleic Acids Res 2006;34:W3–5.
[4] http://www.pathguide.org/.
[5] Stein LD. Integrating biological databases. Nat Rev Genet 2003;4:337–45.
[6] http://www.w3.org/RDF/.
[7] http://www.w3.org/2004/OWL/.
[8] Aduna Sesame, http://www.openrdf.org.
[9] http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid).
[10] Davidson SB, Overton C, Buneman P. Challenges in integrating biological data sources. J Comput Biol 1995;2:557–72.
[11] Köhler J, Philippi S, Lange M. SEMEDA: ontology based semantic integration of biological databases. Bioinformatics 2003;19:2420–7.
[12] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34:D354–7.
[13] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. . Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 2005;33:D428–32.
[14] Life Sciences Identifier, http://www.omg.org/cgi-bin/doc?lifesci/2003-12-02.
[15] Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 2000;16:184–5.
[16] BioPAX: Biological Pathway Exchange, http://www.biopax.org.
[17] The UniProt Consortium. The Universal Protein Resource (UniProt). Nucl. Acids Res 2007; 35: D193-D197.
[18] Cheung KH, Yip KY, Smith A, Deknikker R, Masiar A, Gerstein M. YeastHub: a semantic Web use case for integrating data in the life sciences domain. Bioinformatics 2005;21(1):i85–96.
[19] Shaban-Nejad A, Baker C, Haarslev V, Butler G. The FungalWeb ontology: semantic web challenges in bioinformatics and genomics. The Semantic Web—ISWC 2005, 2005;3729:1063–1066.
[20] http://dev.isb-sib.ch/projects/uniprot-rdf/migration.html.
[21] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/rdf/core.owl.
[22] The Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/.
[23] Baker C, Su X, Butler G, Haarslev V. Ontoligent Interactive Query Tool. Semantic Web Beyond Comput Hum Exper 2006;2:155–69.
[24] http://www.racer-systems.com.
[25] Stephens S, LaVigna D, DiLascio M, Luciano J. Aggregation of bioinformatics data using semantic web technology. J Web Semantic 2006:4.
[26] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33:D514–7.
[27] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 2007;35:D26–31.
[28] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontol Consortium Nat Genet 2000;25:25–9.
[29] Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. Data integration and genomic medicine, J Biomed Inform, 40(1), Bio*Medical Informatics, February 2007, p. 5–16.
[30] http://java.sun.com/products/jsp/jstl/.
[31] http://tuckey.org/urlrewrite/.
[32] Huynh D, Mazzocchi S, Karger D. Piggy bank: experience the semantic web inside your web browser, International Semantic Web Conference (ISWC) 2005.
[33] http://simile.mit.edu/welkin/.
[34] http://lsids.sourceforge.net/resources/firefox-lsid-browser/.
[35] Gruber T. Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud 1995;43:907–28.
[36] http://bio2rdf.org/bio2rdf-2007-02.owl.
[37] The Dublin Core Metadata Initiative, http://dublincore.org/.
[38] Brickley D, Miller L. FOAF Vocabulary Specification, http://xmlns.com/foaf/spec/.
[39] Knouf N. bibTeX Definition in Web Ontology Language (OWL) Version 0.1, Working Draft. http://zeitkunst.org/bibtex/0.1/, 2004.
[40] http://simile.mit.edu/RDFizers/.
[41] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/efetch_help.html.
[42] Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, et al. Ensembl 2007. Nucleic Acids Res 2007;35:D610–7.

[43] Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, et al. The PROSITE database. Nucleic Acids Res 2006;34:D227–30.
[44] Thomas Fielding R. Architectural Styles and Design of Network-based Software Architectures, PhD Thesis, University of California, 2000.
[45] http://www.w3.org/DesignIssues/LinkedData.
[46] Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, et al. Mouse genome database group. The mouse genome database (MGD): from genes to mice—a community resource for mouse biology. Nucleic Acids Res 2005;33:D471–5.
[47] HUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK, http://www.genenames.org/.
[48] http://www.berkeleybop.org/ontologies/.
[49] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[50] Degtyarenko K, Matos PD, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI—Chemical Entities of Biological Interest. Nucleic Acids Res, Database Summary Paper 646.
[51] van Assem M, Malaisé V, Miles A, Schreiber G. A method to convert thesauri to SKOS. Semantic Web Res Appl 2006:95–109.
[52] NCBI User system requirements, http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help. html#UserSystemRequirements.
[53] http://beta.uniprot.org.
[54] Sahoo SS, Bodenreider O, Zeng K, Sheth A. An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information, http://www2007.org/workshops/paper_149.pdf.
[55] Berners-Lee T, Chen Y, Chilton L, Connolly D, Dhanaraj R, Hollenbach J, et al. Tabulator: exploring and analyzing linked data on the semantic web. In: Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06) workshop, Athens, Georgia, 6 November 2006.
[56] http://dig.csail.mit.edu/2007/tab/tabtutorial.html.
[57] http://bio2rdf.org/authority.
[58] http://www.w3.org/TR/rdf-sparql-query/.
[59] Good BM, Wilkinson MD. The life sciences semantic web is full of creeps! Brief Bioinform 2006;7:275–86.
[60] BioRDF subgroup of the HCLS community, http://www.w3.org/2001/sw/hcls/.