Bio2RDF: Towards a Mashup to build bioinformatics knowledge systems

FRANCOIS BELLEAU , MARC-ALEXANDRE NOLIN NICOLE TOURIGNY , PHILIPPE RIGAULT , JEAN MORISSETTE

Integrating Data across web

- Two ways of looking for genomic information
  - Google It !!!
  - Specialized tools like NCBI Entrez
- What about the other databases ???
  - Every year new list of bioinformatic database is available
  - Data integration difficult by traditional data warehouses

Who bells the Cat ????

- W3C !!
  - Proposed a solution based on a series of standards
  - RDF for document and OWL for ontology
  - RDF and OWL generate a triple – subject,predicate and object
  - Database systems capable of handling triples are known as triplestore

Bio2RDF – A Mashup

- Integrates data from more than one source
- Integrates data from popular public databases
- Bio2RDF is a semantic web approach for data integration

Integration using Semantic approach

- Describing and building knowledge systems for semantic web is a challenge for bioinformatic community
- A few specialized projects like YeastHub and FungalWeb are successful to a certain extent
- Bio2RDF is an attempt in this area to integrate data from different sources

Materials and Methods

- Two main ideas of development
  - Conversion of existing databases into RDF format
  - Use semantic web software to merge, query and visualize data
  - Protégé ontology editor,Piggy Bank,Welkin and LSID browser

- Ontology Design
  - Ontology by definition is explicit specification of conceptualization
  - Analyze existing HTML pages, identify predicates and relations describing the entities
  - A hyperlink corresponds to a URI and a label to its predicate
  - OWL description for each selected HTML document created.

RDFizing

- RDFizer were necessary for two key objectives
  - Mapping between data elements of the original document and the predicates in RDF version
  - Normalization of URI according to Bio2RDF syntax
- RDFizer programs for Bio2RDF written in JSP
- Three kinds of RDFizing carried out
  - XML to RDF
  - SQL to RDF
  - Text to RDF

URI Normalization

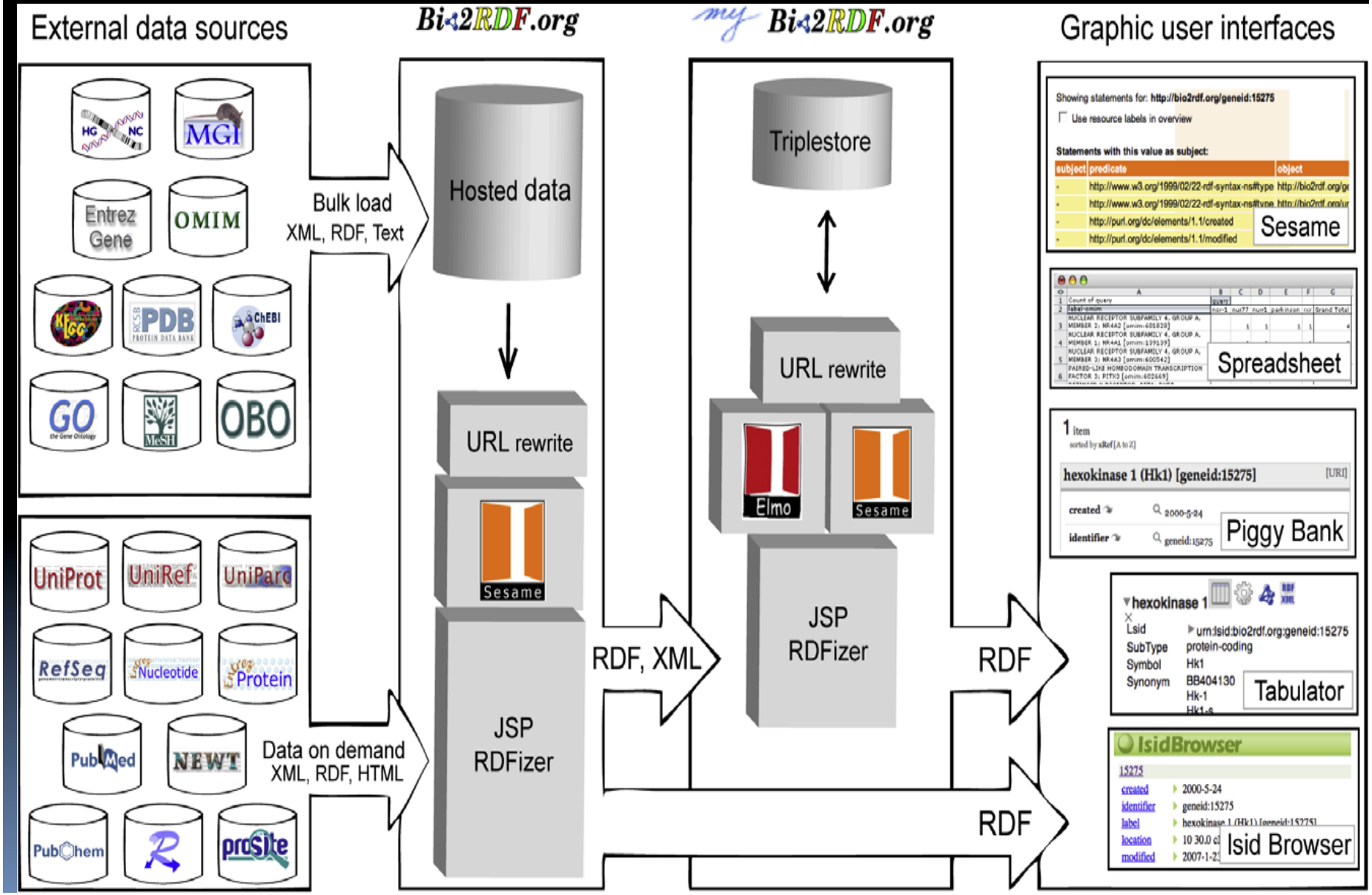
- Normalized URIs needed to allow proper connection of triples
- No links would be created if more than one way of expressing URI existed.
  - http://www.geneontology.org/go#GO:0004396
  - http://purl.uniprot.org/go/0004396
  - urn:lsid:geneontology.org.lsid.biopathways.org:go:0004396
  - All the above represent the same hexokinase, but they are not linked since their URIs are different

A solution in Bio2RDF

- The Strategy
  - Use a REST like interface
  - Lowercase all the URI up to the colon
  - All URIs should return an RDF document
- Syntax of a Normalized Bio2RDF URI
  - http://bio2rdf.org/<namespace>:<identifier>

- Representational State Transfer enables us to produce a stable and clear URI for every document
- The URI case sensitivity poses a problem because each different case results in a theoretically different URI
- If URI for a document creates web page instead of RDF, Linking of data difficult

Bio2RDF Architecture



ELMO Crawler and SESAME Interface

- Elmo crawls RDF documents from the Bio2RDF website
- Sesame interface allows users to browse and query the knowledge base with SeRQL

Three Specific Services added to allow ELMO crawl Specific Knowledge

- To obtain a list of URIs corresponding to the results of a text search using the search engine of the corresponding website.
- To request all URIs in the triplestore which belongs to the specified namespace.
- To create a synonym node to link two URIs which have the same id but different synonymous namespaces.

Results of 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		

Parkinson's Use Case

- An intro to Parkinson's – A slow progressive neurodegenerative disorder
- Four genes Rxr,Nurr1,Nur77 and Nor-1 are of interest in parkinson's
- Major questions that can be answered by Bio2RDF
  - Which GO terms describe our four genes of interest (Rxr, Nurr1, Nur77, and Nor-1)?
  - 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?

A Simple query to find the GO terms!!

```
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}
```

Query to find annotations of cytoplasm and apoptosis

```
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    )
```

Compatibility with ongoing semantic web projects

- Bio2RDF compatible with ongoing semantic web projects
- Compatible with tabulator and various LSID browsers
- The RDF graph returned by Bio2RDF makes it compatible with facet browsers like piggy bank

Extendability and Scalability

- Simple steps to add new database sources
  - Design RDF document to represent data
  - Write corresponding rdfizer programs
  - Install new rdfizer under Bio2RDF servlet of the myBio2RDF installation
  - Add a rewrite rule to the urlrewrite.xml configuration file
  - Restart the myBio2RDF servlet

A Work in Progress..

- The ontology and rdfizer are not definitive
- The ontology still in early stages of development
- The project is open source and can be accessed at bio2rdf.sourceforge.net

THANK YOU