



Integrating Life Sciences Data on the Web using SPARQL

Lee Feigenbaum
May, 2006

SPARQL is...

- ...a query language for selecting values from RDF graphs
- ...a protocol for issuing queries via HTTP GET, HTTP POST, or SOAP
- ...a W3C Candidate Recommendation
- ...capable of returning results serialized as web-friendly JSON structures
- ...perfect for mashing up disparate data sources representable as RDF

```
PREFIX foaf: <...foaf/0.1/>
PREFIX rdf: <...22-rdf-syntax-ns#>
SELECT ?name ?email
WHERE {
    ?person rdf:type foaf:Person .
    ?person foaf:name ?name .
    OPTIONAL {
        ?person foaf:mbox ?email .
    }
}
```

?name	?email
Lee Feigenbaum	feigenbl@us.ibm.com
Grandma Feigenbaum	<i>(unbound)</i>

The Scenario

- Provide a simple, one-stop answer to the question:

How can I discover proteins that are relevant to my work and locate antibodies that target those proteins?

The Data Sources

- Entrez protein sequence and gene databases
 - National Center for Biotechnology Information (NCBI)
 - <http://www.ncbi.nlm.nih.gov/>
 - RDF ← LSID metadata
- Antibody directory
 - Alzheimer Research Forum (AlzForum)
 - <http://alzforum.org/res/com/ant/default.asp>
 - RDF ← HTML scraping
- Mapping data between genes and antibodies
 - Alan Ruttenberg, Millennium
 - RDF ← spreadsheet data
- Taxonomy information
 - Wikispecies, free species directory
 - <http://www.wikispecies.org>
 - RDF ← XSLT applied to XHTML

The Tools

- JavaScript SPARQL client library
 - Issue SPARQL SELECT queries and retrieve results as JavaScript objects
 - Supports all SPARQL endpoints returning JSON results (SPARQLer, Rasqal, XMLArmyKnife, ...)
 - <http://www.thefigtrees.net/lee/sw/sparql.js>

- JSON
 - Lightweight serialization of data structures (e.g. SPARQL resultsets)
 - <http://www.json.org>

- Microtemplates
 - Automagically bind JavaScript-object data to DHTML fragments
 - <http://www.microtemplates.org>

The Demo

Antibodies RDF Demo

The demo's purpose is to demonstrate the power of [SPARQL](#) against distributed life-sciences data sources on the web. This demo's scenario revolves around a researcher searching the NCBI's Entrez Protein database, identifying a protein of interest from the returned results, and then searching for antibodies against that target protein. This demo uses SPARQL to query over these data sources:

- [Entrez Protein](#)
- [Alzheimer Research Forum Antibody Database](#)
- [Wikispecies directory of species](#)

Search input:

<p>NP_003912 (NCBI)</p> <p>B-cell CLL/lymphoma 10</p> <p>Homo sapiens</p>	<p>Bcl-10 (AlzForum)</p> <p>Distributor: BD Pharmingen (cat. no. 551340)</p> <p>Immunogen:</p> <p>Specificity: 31 kDa Bcl-10</p>
<p>NP_776216 (NCBI)</p> <p>mucosa associated lymphoid tissue lymphoma translocation protein 1 isoform b</p> <p>Homo sapiens</p>	<p>Bcl-10 (AlzForum)</p> <p>Distributor: exalpha Biologicals (cat. no. X1119P)</p> <p>Immunogen: synthetic peptide corr. to aa. 5-19 of human bcl-10, N-term</p> <p>Specificity: Bcl-10</p>
<p>NP_006776 (NCBI)</p> <p>mucosa associated lymphoid tissue lymphoma translocation protein 1 isoform a</p> <p>Homo sapiens</p>	<p>Bcl-10 (AlzForum)</p> <p>Distributor: Abcam (cat. no. AB1142)</p> <p>Immunogen: immunogen = synthetic peptide: EMFLPLRS RTVSRQC, human</p> <p>Specificity: Reacts with the C terminal sequence [EMFLPLRS RTVSRQC] of Bcl-10</p>

Done Adblock

What We Learned

Take-away Lessons

- *“With a query language, a client can design their own interface.”*
- Leigh Dodds
- SPARQL + JSON is a powerful Web 2.0 environment
- Even data sources not natively expressed in RDF can be mashed up with SPARQL
- Life sciences provides a rich domain of situational problems to approach with SPARQL-based mashups

Looking Ahead

- As we deal in larger and larger data sets, on-the-fly RDF creation becomes impractical, so:
 - “Smart” federation
 - Dedicated SPARQL endpoints
- Universal naming, merged graphs, and shared predicates only get us so far, so:
 - Custom relations
 - owl:sameAs
 - Human-guided curation

Next Steps

- More data sources!
 - Antibody distributors' databases (price, etc.)
 - Antibodies not related to neuroscience, and for other species
- Integration with NCBI website (e.g. GreaseMonkey script)
- Generate authoritative RDF data via GRDDL transformations or RDFa

Thanks!

- Questions?
- More information: feigenbl@us.ibm.com
- Demo online at <http://thefigtrees.net/lee/sw/demos/antibodies/>
- Thanks to:
 - Alan Ruttenberg, Millennium
 - June Kinoshita and Colin Knep, Alzheimer Research Forum
 - Elias Torres, Ben Szekely, and Alister Lewis-Bowen, IBM