

Combining Public and Proprietary Data

(with semantic web technology)


Bob DuCharme

Washington DC Semantic Web Meetup
March 26, 2010



Introductions

- Presentation and all its URLs:
<http://www.snee.com/semwebmeetup/march26>
- Me: SGML and XML at Moody's, LexisNexis, Innodata Isogen, TopQuadrant
- Weblog: <http://www.snee.com/bobdc.blog>



Outline

- Demo app: what
- The Semantic Web, RDF, and Linked Data
- Demo app
 - overall architecture
 - how, with open source software
 - how, with TopQuadrant products
- A little more about TopQuadrant
- Questions


© Copyright 2007-2010 TopQuadrant Inc. Slide 3



Demo Application

Goal: enhance (fake) analyst buy/sell/hold recommendations with other data about companies to make a more informative, attractive report.


© Copyright 2007-2010 TopQuadrant Inc. Slide 4



Data sources

- analystRecs.xls: analyst recommendations
- finance.yahoo.com: current trading info
- Wikipedia: company summary information
- (SQL database of customer holdings)


© Copyright 2007-2010 TopQuadrant Inc. Slide 5



The Semantic Web

A set of standards and best practices for sharing data and the semantics of that data over the web for use by applications.

© Copyright 2007-2010 TopQuadrant Inc. Slide 6




The Semantic Web

A **set of standards** and best practices for sharing data and the semantics of that data over the web for use by applications.

- RDF
- (OWL, RDFS)
- SPARQL

© Copyright 2007-2010 TopQuadrant Inc. Slide 7



RDF

- Resource Description Framework
- Store data about anything, but especially metadata about resources
- Stored where?
- Very easily aggregated

An RDF “statement”: the triple

- (Subject, predicate, object)
- “index.html has the title ‘My Home Page’.”
- Easily stores (resource ID, propertyName, propertyValue) assertions

TopQuadrant™



Triples

```
# rdf1.nt: sample RDF file in n-triples format.
```

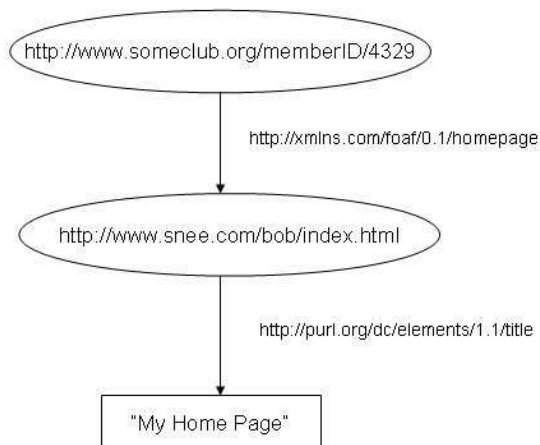
```
<http://www.snee.com/bob/index.html>  
<http://purl.org/dc/elements/1.1/title>  
"My Home Page".
```

```
<http://www.someclub.org/memberID/4329>  
<http://xmlns.com/foaf/0.1/homepage>  
<http://www.snee.com/bob/index.html>.
```

TopQuadrant™



Linking triples into a “graph”



TopQuadrant™



SPARQL

- SPARQL Protocol and RDF Query Language
- Became W3C standard January 2008

TopQuadrant™



SPARQL query 1

```
PREFIX a: <http://www.snee.com/ns/abook#>
SELECT ?s
WHERE { ?s a:firstName "Jim" }
```

Result:

```
-----
| s |
=====
| <http://www.snee.com/ns/id/jimgartner> |
| <http://www.snee.com/ns/id/i129> |
-----
```

TopQuadrant™



SPARQL query 2

```
PREFIX a: <http://www.snee.com/ns/abook#>
SELECT ?ln
WHERE { ?s a:firstName "Jim".
        ?s a:lastName ?ln.
}
```

Result:

```
-----
| ln |
=====
| "Gartner" |
| "Gabriel" |
-----
```

TopQuadrant™



SPARQL query 3

```
PREFIX a: <http://www.snee.com/ns/abook#>
SELECT ?fn ?ln
WHERE { ?s a:firstName ?fn;
          a:lastName ?ln;
          a:instrument "guitar".
}
```

Result:

fn	ln
"Jason"	"Lyman"
"Jaye"	"Urgo"



Is SPARQL Difficult?

The screenshot shows a web browser window with the address bar displaying a URL from cnm.com. The page is titled "Making sense of the 'semantic Web'" and is dated December 15, 2008. The article is by Steve Moliman for CNN. The main text discusses the challenges of the Semantic Web, noting that while it's not as easy as it sounds, it could make surfing Web 3.0 more rewarding. It also mentions that a pair of German researchers have created an experimental kiosk for using semantic web capabilities.



Is SPARQL Difficult?

“Consider, for instance, SPARQL, a query language. To find, say, music artists associated with the producer Timbaland, you'd have to type a long piece of convoluted code that most of us wouldn't bother to do.”

© Copyright 2007-2010 TopQuadrant Inc.

Slide 17



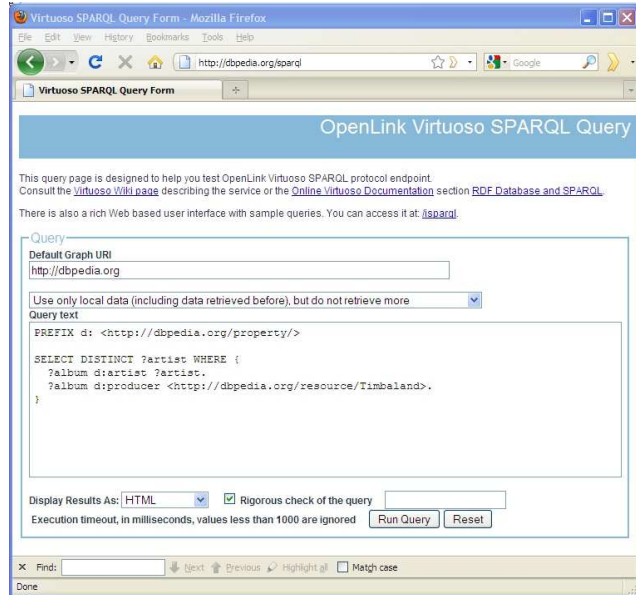
This query...

```
PREFIX d: <http://dbpedia.org/property/>
SELECT DISTINCT ?artist WHERE {
    ?album d:artist ?artist.
    ?album d:producer
        <http://dbpedia.org/resource/Timbaland>.
}
```

© Copyright 2007-2010 TopQuadrant Inc.

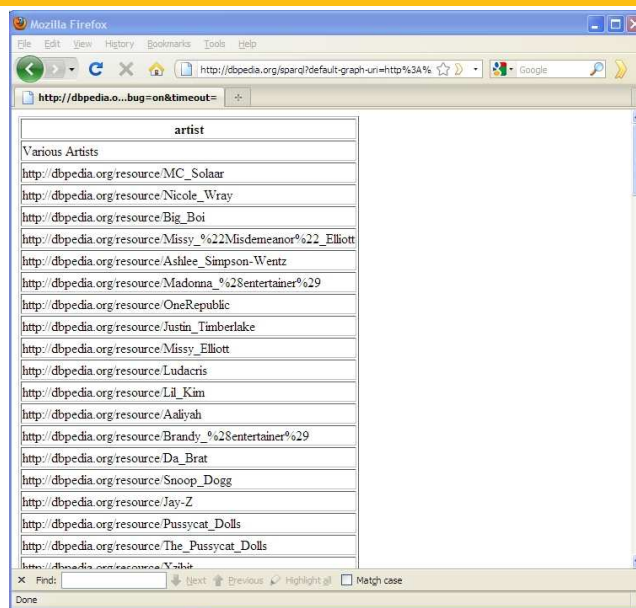
Slide 18

entered here...




© Copyright 2007-2010 TopQuadrant Inc. Slide 19

and there they are.



artist
Various Artists
http://dbpedia.org/resource/MC_Solaar
http://dbpedia.org/resource/Nicole_Wray
http://dbpedia.org/resource/Big_Boi
http://dbpedia.org/resource/Missy_%22Misdemeanor%22_Elliott
http://dbpedia.org/resource/Ashlee_Simpson-Wentz
http://dbpedia.org/resource/Madonna_%28entertainer%29
http://dbpedia.org/resource/OneRepublic
http://dbpedia.org/resource/Justin_Timberlake
http://dbpedia.org/resource/Missy_Elliott
http://dbpedia.org/resource/Ludacris
http://dbpedia.org/resource/Lil_Kim
http://dbpedia.org/resource/Aaliyah
http://dbpedia.org/resource/Brandy_%28entertainer%29
http://dbpedia.org/resource/Da_Brat
http://dbpedia.org/resource/Snoop_Dogg
http://dbpedia.org/resource/Jay-Z
http://dbpedia.org/resource/Pussycat_Dolls
http://dbpedia.org/resource/The_Pussycat_Dolls
http://dbpedia.org/resource/Ya-Ya

© Copyright 2007-2010 TopQuadrant Inc. Slide 20



On the other hand...


Some JavaScript from a View Source of that same CNN page:

```

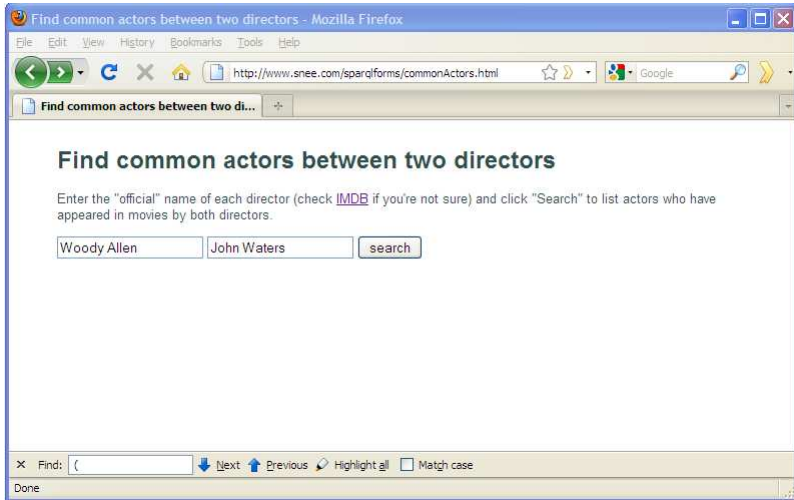
if(cnnWinExtraRegExp.test(cnnWinExtra)){var cnnOmniExtra =
cnnWinExtraRegExp.split(cnnWinExtra);cnnWinLoc = cnnWinLoc +
cnnOmniExtra[0];} else {cnnWinLoc = cnnWinLoc +
cnnWinExtra;}} if (typeof(cnnPageName) != "undefined")
{s.pageName = cnnPageName;s.eVar1 = cnnPageName;} else
{s.pageName = cnnWinLoc;s.eVar1 = cnnWinLoc;} if
(typeof(cnnSectionName) != "undefined")
{s.channel=cnnSectionName;s.eVar2=cnnSectionName;} else
{s.channel="Nonlabeled";s.eVar2="Nonlabeled";} if
(typeof(cnnSubSectionName) != "undefined")
{s.server=cnnSubSectionName;s.eVar3=cnnSubSectionName;} else
{s.server="";s.eVar3="";} if (typeof(cnnSectionFront) !=
"undefined") {s.prop1=cnnSectionFront;} if
(typeof(cnnContentType) != "undefined")
{s.prop4=cnnContentType;s.prop6=s.pageName;}

```


© Copyright 2007-2010 TopQuadrant Inc.
Slide 21

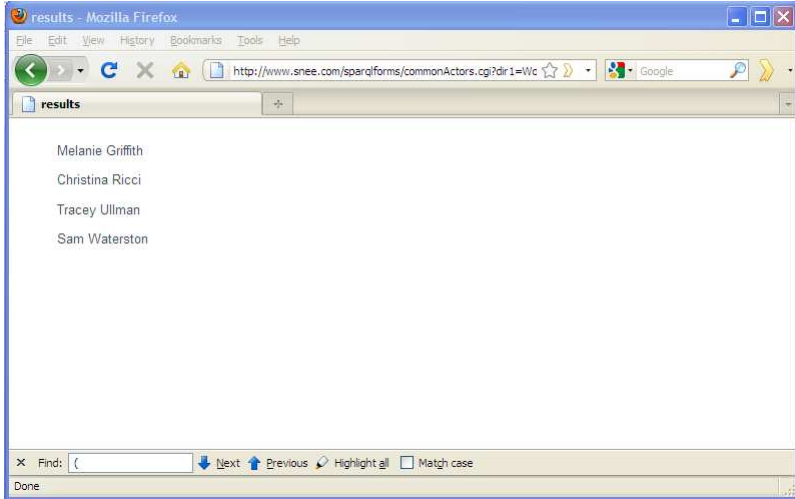


Form-based SPARQL app



© Copyright 2007-2010 TopQuadrant Inc.
Slide 22

 **Form-based SPARQL app: results**



results - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.snee.com/sparqlforms/commonActors.cgi?dir1=Wc

results


Melanie Griffith
Christina Ricci
Tracey Ullman
Sam Waterston

Find: Next Previous Highlight all Match case

Done

© Copyright 2007-2010 TopQuadrant Inc.


Slide 23

 **SPARQL's role in today's message**

- SPARQL lets you query a set of triples
- Data from different public and private sources in different formats can be treated as triples
- Different sets of triples can be easily combined
- So SPARQL lets you mix and match and query data from different sources

© Copyright 2007-2010 TopQuadrant Inc.


Slide 24



The Semantic Web

A set of standards and **best practices** for sharing data and the semantics of that data over the web **for use by applications.**

© Copyright 2007-2010 TopQuadrant Inc. Slide 25



Linked Data

W3C wiki: "LinkedData is to spreadsheets and databases what the Web of hypertext documents is to word processor files."

Jim Hendler: "My document can point at your document on the Web, but my database can't point at something in your database without writing special purpose code. The Semantic Web aims at fixing that."

Kingsley Idehen: "It's a deliverable from the "Semantic Web Project". It adds reference & access granularity to existing #web."

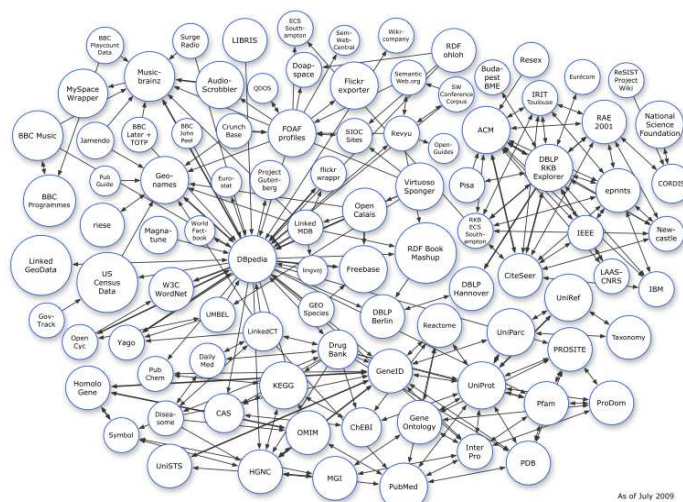
Me: The semantic web without the semantics.


Tim Berners-Lee's Four Linked Data Principles

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.



Linked Data Cloud (August 2009)





Accessing data as RDF

- How do we get RDF out of these data sources so that we can use SPARQL to manipulate that data?
 - Wikipedia
 - analystRecs.xls
 - finance.yahoo.com
 - (SQL database: D2RQ)

© Copyright 2007–2010 TopQuadrant Inc.
Slide 29



Wikipedia infoboxes



© Copyright 2007–2010 TopQuadrant Inc.
Slide 30

Everything Bart wrote on blackboard in season 12

```
SELECT ?episode,?chalkboard_gag
WHERE { ?episode skos:subject
<http://dbpedia.org/resource/Category:The_Simpsons_episodes%2C_season_12>.
?episode dbpedia2:blackboard ?chalkboard_gag }
```

SPARQL results:

episode	chalkboard_gag
A_Tale_of_Two_Springfields	"I will not plant subliminal" messa" gore"s""@en
Bye_Bye_Nerdie	"I will not scare the vice president""@en
Children_of_a_Lesser_Clod	"Today is not Mothra's Day""@en
Day_of_the_Jackanapes	"The hamster did not have a 'full life'""@en
HOMR	Network_television
Homer_vs_Dignity	"I will not surprise the incontinent""@en
Hungry%2C_Hungry_Homer	"Temptation Island was not a sleazy piece of crap""@en
%27m_Goin%27_to_Praiseland	"Genetics is not an excuse""@en
Insane_Clown_Poppy	"I will not surprise the incontinent""@en
Lisa_the_Tree_Hugger	"I am not the acting President.""@en
New_Kids_on_the_Blecch	"I will not buy a presidential pardon""@en
Pokey_Mom	Who_Let_the_Dogs_Out%3F
Simpson_Safari	"I will not flush evidence""@en
Simpsons_Tall_Tales	"I should not be twenty-one by now""@en
Skinner%27s_Sense_of_Snow	"Science class should not end in tragedy""@en
Tennis_the_Menace	"I will not publish the principal's credit report""@en
The_Computer_Wore_Menace_Shoes	"I will only provide a urine sample when asked""@en
The_Great_Money_Caper	"The nurse is not dealing""@en
Trilogy_of_Error	"Fire is not the cleanser""@en



Retrieving DBpedia data

- A query like this

```
CONSTRUCT { <http://dbpedia.org/resource/IBM> ?p ?o }
WHERE { <http://dbpedia.org/resource/IBM> ?p ?o }
```

- Can be stored in a URL like this:

```
http://dbpedia.org/sparql?default-graph-
uri=http%3A%2F%2Fdbpedia.org&query=CONSTRUCT%20%7B%20
%3Chttp%3A%2F%2Fdbpedia.org%2Fresource%2FGoogle%3E%20
%3Fp%20%3Fo%20%7D%20WHERE%20%7B%20%3Chttp%3A%2F%2Fdbp
edia.org%2Fresource%2FGoogle%3E%20%3Fp%20%3Fo%20%7D
```




Accessing spreadsheets as RDF

- RDF123
- XLWrap (server-based; doc includes list of others, including TopBraid Composer)
- My own Perl script

© Copyright 2007-2010 TopQuadrant Inc.

Slide 33



Spreadsheet as CSV (1st 4 lines)

```
analyst,Ticker Symbol,Wikipedia
ID,recommendation,date-time,description
Nick Perkins,GOOG,Google,SELL,2009-12-
14T13:36:00,"Google has had ...
Liz Ford,VOD,Vodafone,BUY,2009-12-
15T18:24:00,"Vodafone has had ...
Betty Bailey,SNE,Sony,HOLD,2009-12-
16T17:21:00,"Sony has had ...
```

© Copyright 2007-2010 TopQuadrant Inc.

Slide 34

TopQuadrant™

finance.yahoo.com

© Copyright 2007-2010 TopQuadrant Inc. Slide 35

TopQuadrant™

Accessing CSV stock ticker data

`http://download.finance.yahoo.com/d/quotes.csv?f=s11d1t1ohgv&e=.csv&s=BUD,IBM,SNE`


returns this:

```
"BUD",47.88,"1/21/2010","4:00pm",49.02,
49.12,47.77,986334

"IBM",129.00,"1/21/2010","4:00pm",130.4
7,130.69,128.06,9608596

"SNE",34.36,"1/21/2010","4:02pm",34.75,
34.95,34.06,1575733
```

© Copyright 2007-2010 TopQuadrant Inc. Slide 36




Accessing RDF stock ticker data

`http://www.rdfdata.org/cgi/stockquotes.cgi`
`?symbols=BUD,IBM,SNE`

returns this...

© Copyright 2007-2010 TopQuadrant Inc. Slide 37



Accessing RDF stock ticker data

```
<rdf:RDF
  xmlns:sq='http://www.rdfdata.org/2009/12/stockquotes#'
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'>


  <sq:Quote rdf:about='#BUD2010-01-21T16:00:00'>
    <sq:tickerSymbol>BUD</sq:tickerSymbol>
    <sq:lastPrice>47.88</sq:lastPrice>
    <sq:dateTime>2010-01-21T16:00:00</sq:dateTime>
    <sq:openingPrice>49.02</sq:openingPrice>

    <sq:dayHigh>49.12</sq:dayHigh>
    <sq:dayLow>47.77</sq:dayLow>
    <sq:volume>986334</sq:volume>
  </sq:Quote>

  <!-- etc.for IBM and Sony -->

</rdf:RDF>
```

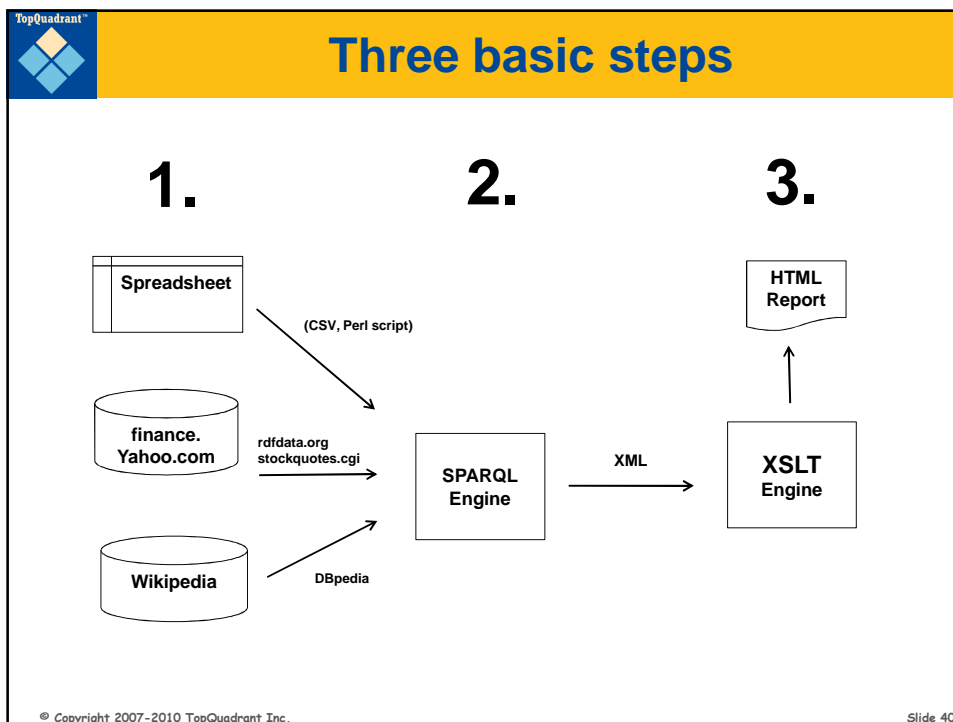
© Copyright 2007-2010 TopQuadrant Inc. Slide 38




Three basic steps

- Combine RDF of analyst recommendations + ticker info + DBpedia company info (+ customer holdings)
- Use SPARQL to mix and match and connect and sort, output XML of result
- Use XSLT to create HTML of report(s)

© Copyright 2007-2010 TopQuadrant Inc.
Slide 39






Jena

- Open source semantic web “framework”
- Java code and utilities
- Supports OWL
- Many useful extensions to SPARQL
- TopBraid products built on it
- Began at Hewlett Packard


© Copyright 2007-2010 TopQuadrant Inc. Slide 41



Tools used by build file

- ARQ: Jena command-line SPARQL tool
- xsltproc: libxml XSLT processor
- Rapper: Redlands utility to convert n3 to RDF/XML
- curl (e.g. curl http://www.google.com > g.txt)
- Perl

© Copyright 2007-2010 TopQuadrant Inc. Slide 42



Running it (part 1 of 2)

```

REM set up classpath for arq
call %ARQROOT%\bat\make_classpath.bat %ARQROOT%


REM Extract spreadsheet data as RDF to analystRecs.rdf.
perl csv2rdf.pl analystRecs.csv > analystRecs.rdf

REM Read analystRecs.rdf, create a one-line script
REM to retrieve RDF of quote data.
xsltproc -o getTickerInfo.bat MakeGetTickerInfo.xsl analystRecs.rdf

REM Get the quote data and put output into quoteData.rdf.
call getTickerInfo

REM Read analystRecs.rdf, create the script to
REM retrieve data from DBpedia
xsltproc -o getDbpediaData.bat MakeGetDbpediaData.xsl analystRecs.rdf
  
```

© Copyright 2007-2010 TopQuadrant Inc. Slide 43



Running it (part 2 of 2)

```


REM run that script, put output in dbpedia.n3
call getDbpediaData

REM convert DBpedia data from n3 to RDF/XML format
rapper -q -i n3 -o rdfxml dbpedia.n3 > dbpedia.rdf

REM extract report data from the combination of the
REM three files with SPARQL query
java -cp %CP% arq.arq --results=XML --query=PickDataForReport.sparql --
  data=analystRecs.rdf --data=quoteData.rdf --data=dbpedia.rdf >
  reportData.xml

REM Create HTML report from report data
xsltproc SPARQLXML2HTML.xsl reportData.xml
  
```

© Copyright 2007-2010 TopQuadrant Inc. Slide 44



Picking the data we need (pt. 1 of 2)


```

PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX xs: <http://www.w3.org/2001/XMLSchema#>
PREFIX ar: <file:///Bob%20sandbox/demo/financial/analystRecs.xls#>
PREFIX sq: <http://www.rdfdata.org/2009/12/stockquotes#>

SELECT ?tickerSymbol ?coName ?analyst ?description ?recommendation
       ?recDateTime ?lastPrice ?quoteDateTime ?dayHigh ?dayLow
       ?openingPrice ?volume ?logo ?revenue ?netIncome ?abstract
       ?thumbnail
WHERE {
    ?analystData ar:analyst ?analyst ;
                ar:description ?description ;
                ar:tickerSymbol ?tickerSymbol ;
                ar:wikipediaID ?wikipediaID ;
                ar:recommendation ?recommendation ;
                ar:dateTime ?recDateTime .

    LET (?coName := ?wikipediaID) .
  
```

© Copyright 2007-2010 TopQuadrant Inc.
Slide 45



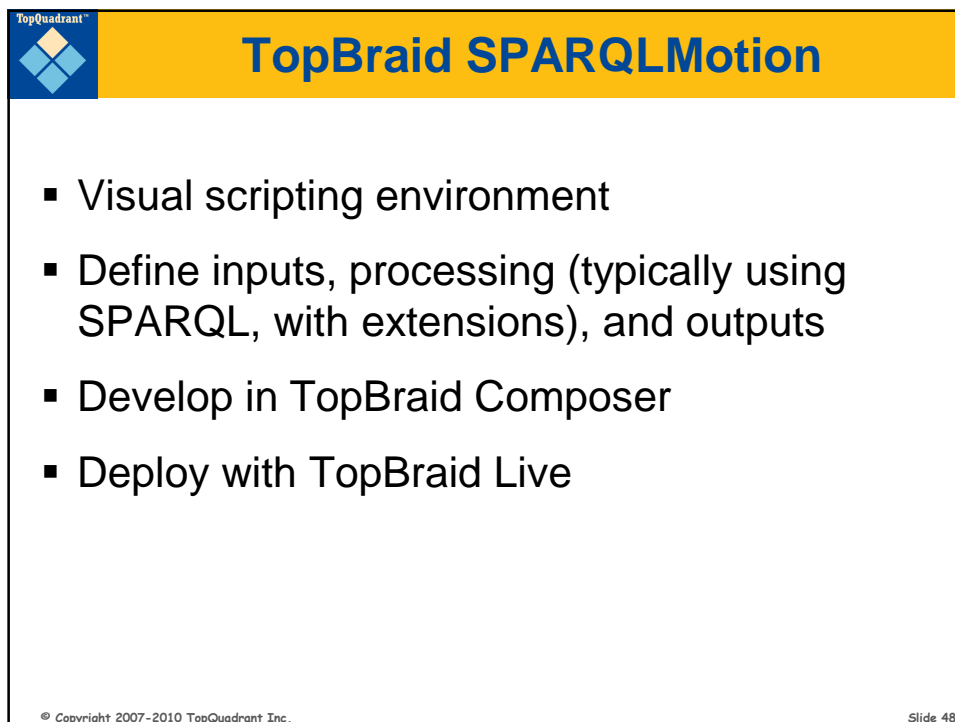
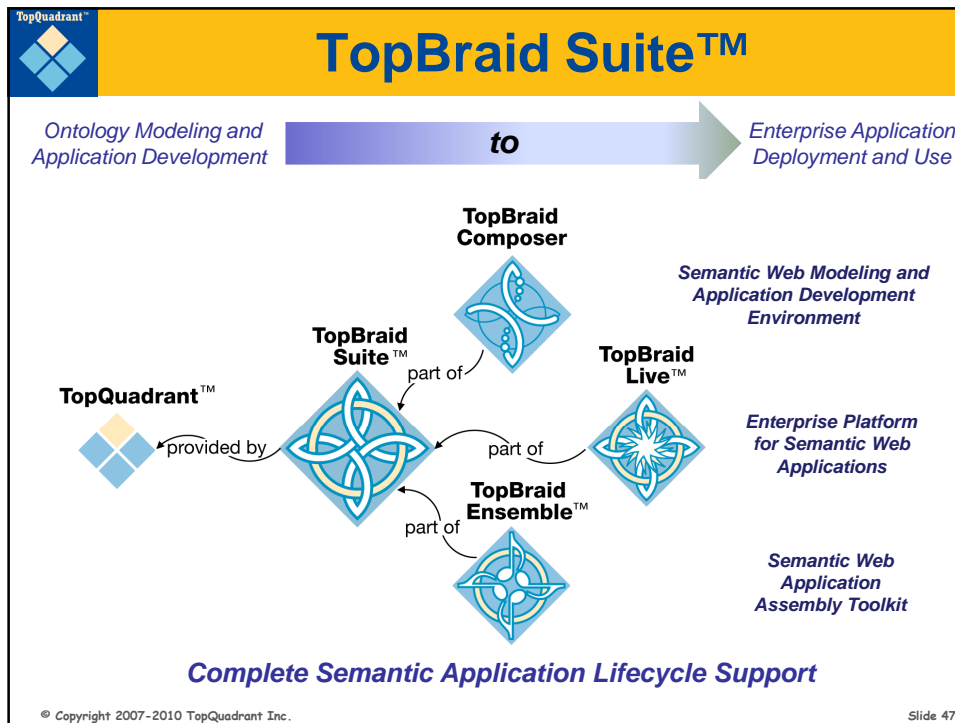
Picking the data we need (pt. 2 of 2)

```

?quoteData sq:tickerSymbol ?tickerSymbol ;
           sq:lastPrice ?lastPrice ;
           sq:dateTime ?quoteDateTime ;
           sq:dayHigh ?dayHigh ;
           sq:dayLow ?dayLow ;
           sq:lastPrice ?lastPrice ;
           sq:openingPrice ?openingPrice ;
           sq:volume ?volume .

?dbpURI <http://dbpedia.org/ontology/revenue> ?revenue .
OPTIONAL {?dbpURI <http://dbpedia.org/property/companyLogo> ?logo . } .
OPTIONAL {?dbpURI <http://dbpedia.org/ontology/netIncome> ?netIncome .} .
OPTIONAL {
    ?dbpURI <http://dbpedia.org/property/abstract> ?abstract .
    FILTER (lang(?abstract) = "en") .
} .
OPTIONAL {?dbpURI <http://dbpedia.org/ontology/thumbnail> ?thumbnail . } .
FILTER regex(fn:substring(xs:string(?dbpURI), 29), ?wikipediaID)
}
  
```

© Copyright 2007-2010 TopQuadrant Inc.
Slide 46





TopQuadrant

- **Formed in 2001**
 - Privately held
 - First Semantic Web Consulting Firm in the U.S.
- **Products: TopBraid Suite**
 - Semantic Web Application Development Platform
- **Solution Services**
 - Jumpstarts to Large Implementations
 - Envisioning Workshop
- **Semantic Web Training**
 - 600+ people Trained
 - On-site and Public Training
- **Locations**
 - Alexandria, Virginia
 - Mountain View, California
 - TopQuadrant Korea – Seoul, S. Korea
- **Strategic Partnerships**
 - Oracle, Franz, CTG





© Copyright 2007-2010 TopQuadrant Inc.

Slide 49



600+ Customers


National Aeronautics
and Space Administration






















































© Copyright 2007-2010 TopQuadrant Inc.


Slide 50



Trying the Software

- 30-day evaluation copy lets you do everything, including all SPARQLMotion application development
- Free version limited to ontology and data editing, never expires
- Both available for Windows, Mac, and Linux

© Copyright 2007-2010 TopQuadrant Inc. Slide 51



Summary

- RDF makes data easier to combine
- Tools exist to treat many kinds of data formats as RDF, whether stored privately or not
- More and more public data is available as RDF
- After combining, SPARQL is a great way to combine, mix and match data
- You can do this all with free software, but it's a lot easier with TopQuadrant's TopBraid Suite

© Copyright 2007-2010 TopQuadrant Inc. Slide 52