LOD 123: Making the semantic web easier to use

Tim Finin

University of Maryland, Baltimore County

Joint work with Lushan Han, Varish Mulwad, Anupam Joshi
Overview

• Linked Open Data 101
• Two ongoing UMBC dissertations
  – Varish Mulwad, Generating linked data from tables
  – Lushan Han, Querying linked data with a quasi-NL interface
Linked Open Data (LOD)

• Linked **data** is just RDF data, typically just the instances (ABOX), not schema (TBOX)

• RDF data is a graph of triples
  
  – URI URI string
    
    dbr:Barack_Obama dbo:spouse “Michelle Obama”
  
  – URI URI URI
    
    dbr:Barack_Obama dbo:spouse dbpedia:Michelle_Obama

• Best **linked** data practice prefers the $2^{nd}$ pattern, using nodes rather than strings for “entities”

• Liked **open** data is just linked data freely accessible on the Web along with any required ontologies
Semantic Web

Use Semantic Web Technology to publish shared data & knowledge

Semantic web technologies allow machines to share data and knowledge using common web language and protocols.

~ 1997

Semantic Web beginning
Semantic Web => Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

Data is inter-linked to support integration and fusion of knowledge

LOD beginning
Use Semantic Web Technology to publish shared data & knowledge.

Data is interlinked to support integration and fusion of knowledge.
Semantic Web => Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

Data is inter-linked to support integration and fusion of knowledge

... and growing
Use Semantic Web Technology to publish shared data & knowledge

LOD is the new Cyc: a common source of background knowledge

Data is interlinked to support integration and fusion of knowledge

2010...growing faster
Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

LOD is the new Cyc: a common source of background knowledge

Data is interlinked to support integration and fusion of knowledge

2011: 31B facts in 295 datasets interlinked by 504M assertions on ckan.net
Exploiting LOD not (yet) Easy

- Publishing or using LOD data has inherent difficulties for the potential user
  - It’s difficult to explore LOD data and to query it for answers
  - It’s challenging to publish data using appropriate LOD vocabularies & link it to existing data

- Problem: $O(10^4)$ schema terms, $O(10^{11})$ instances

- I’ll describe two ongoing research projects that are addressing these problems
Generating Linked Data by Inferring the Semantics of Tables

Research with Varish Mulwad

http://ebiq.org/j/96
Early work

• Mapping tables to RDF led to early tools
  – D2RQ (2006) relational tables to RDF
  – RDF 123 (2007) spreadsheet to RDF
• And a recent W3C standard
  – R2RML (2012) a W3C recommendation
• But none of these can automatically generate high-quality linked data
  – They don’t link to LOD classes and properties nor recognize entity mentions
### Goal: Table => LOD*

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Position</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Jordan</td>
<td>Chicago</td>
<td>Shooting guard</td>
<td>1.98</td>
</tr>
<tr>
<td>Allen Iverson</td>
<td>Philadelphia</td>
<td>Point guard</td>
<td>1.83</td>
</tr>
<tr>
<td>Yao Ming</td>
<td>Houston</td>
<td>Center</td>
<td>2.29</td>
</tr>
<tr>
<td>Tim Duncan</td>
<td>San Antonio</td>
<td>Power forward</td>
<td>2.11</td>
</tr>
</tbody>
</table>

* DBpedia

http://dbpedia.org/class/yago/NationalBasketballAssociationTeams

http://dbpedia.org/resource/Allen_Iverson

Player height in meters
## Goal: Table => LOD*

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Position</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Jordan</td>
<td>Chicago</td>
<td>Shooting guard</td>
<td>1.98</td>
</tr>
<tr>
<td>Allen Iverson</td>
<td>Philadelphia</td>
<td>Point guard</td>
<td>1.83</td>
</tr>
<tr>
<td>Yao Ming</td>
<td>Houston</td>
<td>Center</td>
<td>2.29</td>
</tr>
<tr>
<td>Tim Duncan</td>
<td>San Antonio</td>
<td>Power forward</td>
<td>2.11</td>
</tr>
</tbody>
</table>

* DBpedia

All this in a completely automated way
Tables are everywhere!! ... yet ...

The web – 154 million high quality relational tables
Evidence-based medicine

Evidence-based medicine judges the efficacy of treatments or tests by meta-analyses of clinical trials. Key information is often found in tables in articles.

Figure: Evidence-Based Medicine - the Essential Role of Systematic Reviews and the Need for Automated Text Mining Tools, IHI 2010.
~ 400,000 datasets 😊😊

~ < 1% in RDF 😞
2010 Preliminary System

T2LD Framework

Predict Class for Columns → Linking the table cells → Identify and Discover relations

Class prediction
Entity Linking

Examples of class labels:
Column – Nationality
Prediction – MilitaryConflict

Column – Birth Place
Prediction – PopulatedPlace
Sources of Errors

• The *sequential* approach let errors percolate from one phase to the next

• The system was biased toward predicting overly general classes over more appropriate specific ones

• *Heuristics* largely drive the system

• Although we consider multiple sources of evidence, we did not *joint assignment*
A Domain Independent Framework

- **Sampling**
- **Acronym detection**

**Pre-processing modules**

- **Query and generate initial mappings**
- **Joint Inference/Assignment**

**Flow**

1. Generate Linked RDF
2. Verify (optional)
3. Store in a knowledge base & publish as LOD

**Table**

<table>
<thead>
<tr>
<th>City</th>
<th>Mayor</th>
<th>State</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>T. Menino</td>
<td>MA</td>
<td>610,000</td>
</tr>
<tr>
<td>New York</td>
<td>M. Bloomberg</td>
<td>NY</td>
<td>8,400,000</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>M. Nutter</td>
<td>PA</td>
<td>1,500,000</td>
</tr>
<tr>
<td>Baltimore</td>
<td>S. Dixon</td>
<td>MD</td>
<td>640,000</td>
</tr>
<tr>
<td>Washington</td>
<td>A. Fenty</td>
<td>DC</td>
<td>595,000</td>
</tr>
</tbody>
</table>
Query Mechanism

Team

| Michael Jordan | Chicago Bulls | Shooting Guard | 1.98 |

{dbo:Place, dbo:City, yago:WomenArtist, yago:LivingPeople, yago:NationalBasketballAssociationTeams...}

possible types

possible entities

Chicago Bulls, Chicago, Judy Chicago ...

.......
Ranking the candidates

• $C_i = "State"$; $L_{ci} = \text{AdministrativeRegion}$

• $f_1 = [\text{Levenshtein distance}(C_i, L_{ci}), \text{Dice Score} (C_i, L_{ci}), \text{Semantic Similarity} (C_i, L_{ci}), \text{InformationGain}(L_{ci})]$}

• $\psi_1 = \exp(w_1^T f_1(C_i, L_{ci}))$
Ranking the candidates

- $R_{ij} = \text{“Baltimore”} ; E_{ij} = \text{Baltimore_Maryland}$

- $f_2 = [\text{Levenshtein distance}(R_{ij}, E_{ij}), \newline \text{Dice Score} (R_{ij}, E_{ij}), \newline \text{PageRank} (E_{ij}), \newline \text{KBScore} (E_{ij}), \newline \text{PageLength} (E_{ij}) ]$

- $\psi_2 = \exp(w_2^T f_2(R_{ij}, E_{ij}))$
Joint Inference over evidence in a table

✓ Probabilistic Graphical Models
A graphical model for tables
Joint inference over evidence in a table

Class

Team

Chicago
Philadelphia
Houston
San Antonio

Instance
Parameterized graphical model

Function that captures the affinity between the column headers and row values

Captures interaction between row values

Captures interaction between column headers

Variable Node: Column header

Row value

Factor Node
Challenge: Interpreting Literals

Many columns have literals, e.g., numbers

- Predict properties based on cell values
- Cyc had hand coded rules: *humans don’t live past 120*
- We extract *value distributions* from LOD resources
  - Differ for subclasses: age of *people vs. political leaders vs. athletes*
  - Represent as *measurements*: value + units
- Metric: possibility/probability of values given distribution
Other Challenges

• Using table *captions* and other text is associated documents to provide context

• **Size** of some data.gov tables (> 400K rows!) makes using full graphical model impractical
  – *Sample* table and run model on the subset

• Achieving acceptable accuracy may require human input
  – 100% accuracy unattainable automatically
  – How best to let humans offer advice and/or correct interpretations?
PMI as an association measure

We use pointwise mutual information (pmi) to measure the association between two RDF resources (nodes)

$$\text{pmi}(x; y) \equiv \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$  

pmi is used for word association by comparing how often two words occur together in text to their expected co-occurrence if independent
PMI for RDF instances

• For text, the co-occurrence context is usually a window of some number of words (e.g, 50)
• For RDF instances, we count three graph patterns as instances of the co-occurrence of N1 and N2

• Other graph patterns can be added, but we’ve not evaluated their utility or cost to compute.
PMI for RDF types

• We also want to measure the association strength between RDF types, e.g., a dbo:Actor associated with a dbo:Film vs. a dbo:Place

• We can also measure the association of an RDF property and types, e.g. dbo:author used with a dbo:Film vs. a dbo:Book

• Such simple statistics can be efficiently computed for large RDF collections in parallel
GoRelations: Intuitive Query System for Linked Data

Research with Lushan Han

http://ebiq.org/j/93
Dbpedia is the Stereotypical LOD

• DBpedia is an important example of Linked Open Data
  – Extracts structured data from Infoboxes in Wikipedia
  – Stores in RDF using custom ontologies Yago terms

• The major integration point for the entire LOD cloud

• Explorable as HTML, but harder to query in SPARQL
Browsing DBpedia’s Mark Twain

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:birthDate</td>
<td>1835-11-30 (xsd:date)</td>
</tr>
<tr>
<td>dbpedia-owl:birthName</td>
<td>Samuel Langhorne Clemens</td>
</tr>
<tr>
<td>dbpedia-owl:birthPlace</td>
<td>dbpedia:Florida, Missouri</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Missouri</td>
</tr>
<tr>
<td>dbpedia-owl:child</td>
<td>dbpedia:Jean_Clemens</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Susy_Clemens</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Clara_Clemens</td>
</tr>
<tr>
<td>dbpedia-owl:deathDate</td>
<td>1910-04-21 (xsd:date)</td>
</tr>
<tr>
<td>dbpedia-owl:deathPlace</td>
<td>dbpedia:Redding, Connecticut</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Connecticut</td>
</tr>
<tr>
<td>dbpedia-owl:genre</td>
<td>dbpedia:Essay</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Historical_fiction</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Social_commentary</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Satire</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Non-fiction</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Philosophy_and_literature</td>
</tr>
<tr>
<td></td>
<td>dbpedia: Literary_criticism</td>
</tr>
<tr>
<td></td>
<td>dbpedia: Travel_literature</td>
</tr>
<tr>
<td></td>
<td>dbpedia: Fiction</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Children’s_literature</td>
</tr>
<tr>
<td>dbpedia-owl:notableWork</td>
<td>dbpedia: Adventures_of_Huckleberry_Finn</td>
</tr>
<tr>
<td>dbpedia-owl:occupation</td>
<td>dbpedia: The_Adventures_of_Tom_Sawyer</td>
</tr>
<tr>
<td>dbpedia-owl:pseudonym</td>
<td>Mark Twain</td>
</tr>
<tr>
<td>dbpedia-owl:thumbnail</td>
<td><a href="http://upload.wikimedia.org/wiki/commons/thumb/c/c5/Mark_Twain_by_AF_Bradley.jpg">http://upload.wikimedia.org/wiki/commons/thumb/c/c5/Mark_Twain_by_AF_Bradley.jpg</a></td>
</tr>
<tr>
<td>dbpedia-owl:pageTitleExternalLink</td>
<td><a href="http://www.sfgate.com/cgi-bin/article.cgi?file=/a/2005/08/19/MNGOBEA9J1.DTL">http://www.sfgate.com/cgi-bin/article.cgi?file=/a/2005/08/19/MNGOBEA9J1.DTL</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.pbs.org/marktwain/scrapbook/index.html">http://www.pbs.org/marktwain/scrapbook/index.html</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.kamakurapens.com/TwainsConklin.html">http://www.kamakurapens.com/TwainsConklin.html</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.ucmerced.edu/faculty/facultybio.asp?facultyid=95">http://www.ucmerced.edu/faculty/facultybio.asp?facultyid=95</a></td>
</tr>
</tbody>
</table>
Querying LOD is Much Harder

• Querying DBpedia requires a lot of a user
  – Understand the RDF model
  – Master SPARQL, a formal query language
  – Understand ontology terms: 320 classes & 1600 properties!
  – Know instance URIs (>1M entities!)
  – Term heterogeneity (Place vs. PopulatedPlace)

• Querying large LOD sets overwhelming

• Natural language query systems still a research goal
Goal

- Allow a user with a basic understanding of RDF to query DBpedia and ultimately distributed LOD collections
  - To explore what data is in the system
  - To get answers to question
  - To create SPARQL queries for reuse or adaptation

- Desiderata
  - Easy to learn and to use
  - Good accuracy (e.g., precision and recall)
  - Fast
Key Idea

Structured keyword queries

Reduce problem complexity by:
– User enters a *simple graph*, and
– Annotates the nodes and arcs with *words and phrases*
Structured Keyword Queries

- Nodes denote entities and links binary relations
- Entities described by two unrestricted terms: *name* or value and *type* or concept
- Result entities marked with ? and those not with *
- A compromise between a natural language Q&A system and SPARQL
  - Users provide compositional structure of the question
  - Free to use their own terms in annotating the structure
Translation – Step One
finding semantically similar ontology terms

For each concept or relation in the graph, generate the $k$ most semantically similar candidate ontology classes or properties.

Similarity metric is distributional similarity, LSA, and WordNet.
Another Example

Football players who were born in the same place as their team’s president
To assemble the best interpretation we rely on statistics of the data.

Primary measure is pointwise mutual information (PMI) between RDF terms in the LOD collection.

This measures the degree to which two RDF terms occur together in the knowledge base.

In a reasonable interpretation, ontology terms associate in the way that their corresponding user terms connect in the structured keyword query.
Translation – Step Two

disambiguation algorithm

Three aspects are combined to derive an overall goodness measure for each candidate interpretation.

Joint disambiguation

Resolving direction

Link reasonableness

\[
\text{argmax}_{G} \text{goodness}(G) = \text{argmax}_{L_i} \sum_{i=1}^{m} \text{goodness}(L_i) \quad (1)
\]

If \([\text{PMI}(c(O_i), p(R_i)) + \text{PMI}(p(R_i), c(S_i))] - [\text{PMI}(c(S_i), p(R_i)) + \text{PMI}(p(R_i), c(O_i))] > \alpha\]

Then \(S'_i = O_i, O'_i = S_i\)

Else \(S'_i = S_i, O'_i = O_i\) \quad (2)

\[
\text{goodness}(L_i) = \max(\text{PMI}(c(S'_i), p(R_i)) \cdot \text{sim}(S'_i, c(S'_i)) \cdot \text{sim}(R_i, p(R_i)) + \text{PMI}(p(R_i), c(O'_i)) \cdot \text{sim}(O'_i, c(O'_i)) \cdot \text{sim}(R_i, p(R_i)), \beta) + \text{PMI}(c(S'_i), c(O'_i)) \cdot \text{sim}(S'_i, c(S'_i)) \cdot \text{sim}(O'_i, c(O'_i)) \quad (3)
\]
Example of Translation result

Concepts: Place => Place, Author => Writer, Book => Book
Properties: born in => birthPlace, wrote => author (inverse direction)
The translation of a semantic graph query to SPARQL is straightforward given the mappings

**Concepts**
- Place => Place
- Author => Writer
- Book => Book

**Relations**
- born in => birthPlace
- wrote => author
Evaluation

• 33 test questions from 2011 *Workshop on Question Answering over Linked Data* answerable using DBpedia

• Three human subjects unfamiliar with DBpedia translated the test questions into semantic graph queries

• Compared with two top natural language QA systems: *PowerAqua* and *True Knowledge*

<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoRelations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>regular</td>
<td>0.687</td>
<td>0.722</td>
<td>0.704</td>
</tr>
<tr>
<td>concise</td>
<td>0.736</td>
<td>0.803</td>
<td>0.768</td>
</tr>
<tr>
<td>PowerAqua</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st triple</td>
<td>0.372</td>
<td>0.483</td>
<td>0.420</td>
</tr>
<tr>
<td>all triples</td>
<td>0.334</td>
<td>0.483</td>
<td>0.395</td>
</tr>
<tr>
<td>merged</td>
<td>0.255</td>
<td>0.291</td>
<td>0.272</td>
</tr>
<tr>
<td>True Knowledge</td>
<td>0.469</td>
<td>0.535</td>
<td>0.500</td>
</tr>
</tbody>
</table>
Please input relations in your query. One per line.

?x/American Football Player, date of birth, ?y/Date  
?x, height, ?z/Number

Examples: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Query  Relaxed Query  See Translations

Message: This query gives data about the height of American football players and their date of birth. Football fans may feel it interesting to know how the height of football players changes over time.

http://ebiq.org/GOR

Powered by DBpedia and Virtuoso
2011 Ebiquity Lab, UMBC.
<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://dbpedia.org/resource/Jack_Shapiro">http://dbpedia.org/resource/Jack_Shapiro</a></td>
<td>1.5494  1907-03-22</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Josh_Betts">http://dbpedia.org/resource/Josh_Betts</a></td>
<td>1.5748  1982-08-25</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Nate_Abrams">http://dbpedia.org/resource/Nate_Abrams</a></td>
<td>1.6256  1897-12-25</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Herbert_Clow">http://dbpedia.org/resource/Herbert_Clow</a></td>
<td>1.6256  1899-05-07</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Earl_Warweg">http://dbpedia.org/resource/Earl_Warweg</a></td>
<td>1.6764  1892-01-11</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Jacquizz_Rodgers">http://dbpedia.org/resource/Jacquizz_Rodgers</a></td>
<td>1.6764  1990-02-06</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/John_Brallier">http://dbpedia.org/resource/John_Brallier</a></td>
<td>1.6764  1876-12-27</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Menz_Lindsey">http://dbpedia.org/resource/Menz_Lindsey</a></td>
<td>1.6764  1897-07-25</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Stefan_Logan">http://dbpedia.org/resource/Stefan_Logan</a></td>
<td>1.6764  1981-06-02</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Cory_Ross">http://dbpedia.org/resource/Cory_Ross</a></td>
<td>1.6764  1982-09-22</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Jack_Fleischman">http://dbpedia.org/resource/Jack_Fleischman</a></td>
<td>1.6764  1901-08-15</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Eddie_Scharer">http://dbpedia.org/resource/Eddie_Scharer</a></td>
<td>1.6764  1902-01-26</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/John_Barrett_American_football">http://dbpedia.org/resource/John_Barrett_American_football</a></td>
<td>1.6764  1899-02-25</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Dick_Thornton_American_football">http://dbpedia.org/resource/Dick_Thornton_American_football</a></td>
<td>1.6764  1908-02-04</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Walter_French_American_baseball">http://dbpedia.org/resource/Walter_French_American_baseball</a></td>
<td>1.7018  1899-07-12</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Al_Loeb">http://dbpedia.org/resource/Al_Loeb</a></td>
<td>1.7018  1890-03-11</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Antwaun_Carter">http://dbpedia.org/resource/Antwaun_Carter</a></td>
<td>1.7018  1981-09-09</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Terry_Caulley">http://dbpedia.org/resource/Terry_Caulley</a></td>
<td>1.7018  1984-06-22</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Kendall_Hunter">http://dbpedia.org/resource/Kendall_Hunter</a></td>
<td>1.7018  1988-09-16</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Mark_McMillian">http://dbpedia.org/resource/Mark_McMillian</a></td>
<td>1.7018  1970-04-29</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Al_Comsweet">http://dbpedia.org/resource/Al_Comsweet</a></td>
<td>1.7018  1906-07-16</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Bodie_Weldon">http://dbpedia.org/resource/Bodie_Weldon</a></td>
<td>1.7018  1895-11-07</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Johnny_Bryan">http://dbpedia.org/resource/Johnny_Bryan</a></td>
<td>1.7018  1897-02-28</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Andrew_Hawkins">http://dbpedia.org/resource/Andrew_Hawkins</a></td>
<td>1.7018  1986-03-10</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Chad_Owens">http://dbpedia.org/resource/Chad_Owens</a></td>
<td>1.7018  1982-04-03</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Mark_Higgs">http://dbpedia.org/resource/Mark_Higgs</a></td>
<td>1.7018  1984-04-11</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Emmett_McLemore">http://dbpedia.org/resource/Emmett_McLemore</a></td>
<td>1.7018  1984-08-17</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Danny_Woodhead">http://dbpedia.org/resource/Danny_Woodhead</a></td>
<td>1.7018  1984-08-17</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Garrett_Wolfe">http://dbpedia.org/resource/Garrett_Wolfe</a></td>
<td>1.7018  1980-02-10</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/resource/Shauld_Williams">http://dbpedia.org/resource/Shauld_Williams</a></td>
<td>1.7018  1980-02-10</td>
</tr>
</tbody>
</table>
Three Top Interpretations of the User Query

Top One Interpretation

Before Step Three

* ?x/americanFootballPlayer has a candidate list including 0:AmericanFootballPlayer 1.00 1:Person 0.25 (the selected choice is AmericanFootballPlayer 1.00)
* ?x/number has a candidate list including 0:Number 1.00 1:Group 0.31 2:Code 0.31 3:EthnicGroup 0.25 4:size 0.23 5:series 0.23 6:single 0.21 7:OlympicResult 0.15 8:MountainRange 0.14 (the selected choice is Number 1.00)
* ?y/date has a candidate list including 0:date 1.00 1:grape 0.19 2:Holiday 0.18 3:name 0.18 4:place 0.15 5:event 0.15 6:era 0.15 7:givenName 0.15 8:HistoricPlace 0.14 9:WrestlingEvent 0.14 10:populatedPlace 0.12 11:sportsEvent 0.12 12:WorldHeritageSite 0.12 13:NaturalPlace 0.12 14:MixedMartialArtsEvent 0.11 (the selected choice is date 1.00)

After Step Three

* ?x/americanFootballPlayer => AmericanFootballPlayer
* ?x/number => Number
* ?y/date => Date
* &?x/americanFootballPlayer date of birth ?y/date => 0/@birthDate 1.00
* &?x/americanFootballPlayer height ?z/number => 0/@height 1.00

The Regular SPARQL Query

PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x, ?z, ?y WHERE {
  ?x a dbo:AmericanFootballPlayer .
  ?x dbo:birthDate ?y .
  ?x dbo:height ?z .
}

The Concise SPARQL Query

PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x, ?z, ?y WHERE {
  ?x a dbo:AmericanFootballPlayer .
  {?x dbo:birthDate ?y} .
  {?x dbo:height ?z} .
}

Top Two Interpretation

Before Step Three

* ?x/americanFootballPlayer has a candidate list including 0:americanFootballPlayer 0.25 (the selected choice is AmericanFootballPlayer 1.00)
* ?x/number has a candidate list including 0:AmericanFootballPlayer 1.00 1:Group 0.31 2:Code 0.31 3:EthnicGroup 0.25 4:size 0.23 5:series 0.23 6:single 0.21 7:OlympicResult 0.15 8:MountainRange 0.14 (the selected choice is AmericanFootballPlayer 1.00)

http://ebiq.org/GOR
Current challenges

• Baseline system works well for DBpedia
• Current challenges we are addressing are
  – Adding direct entity matching
  – Relaxing the need for type information
  – Testing on other LOD collections and extending to a set of distributed LOD collections
  – Developing a better Web interface
  – Allowing user feedback and advice
• See http://ebiq.org/93 for more information & try our alpha version at http://ebiq.org/GOR
Final Conclusions

• Linked Data is an emerging paradigm for sharing structured and semi-structured data
  – Backed by machine-understandable semantics
  – Based on successful Web languages and protocols

• Generating and exploring Linked Data resources can be challenging
  – Schemas are large, too many URIs

• New tools for mapping tables to Linked Data and translating structured natural language queries help reduce the barriers
http://ebiq.org/