LOD for the Rest of Us

Tim Finin, Anupam Joshi,
Varish Mulwad and Lushan Han

University of Maryland,
Baltimore County

15 March 2012

UMBC ebiquity

http://ebiq.org/r/338

TL;DR Version

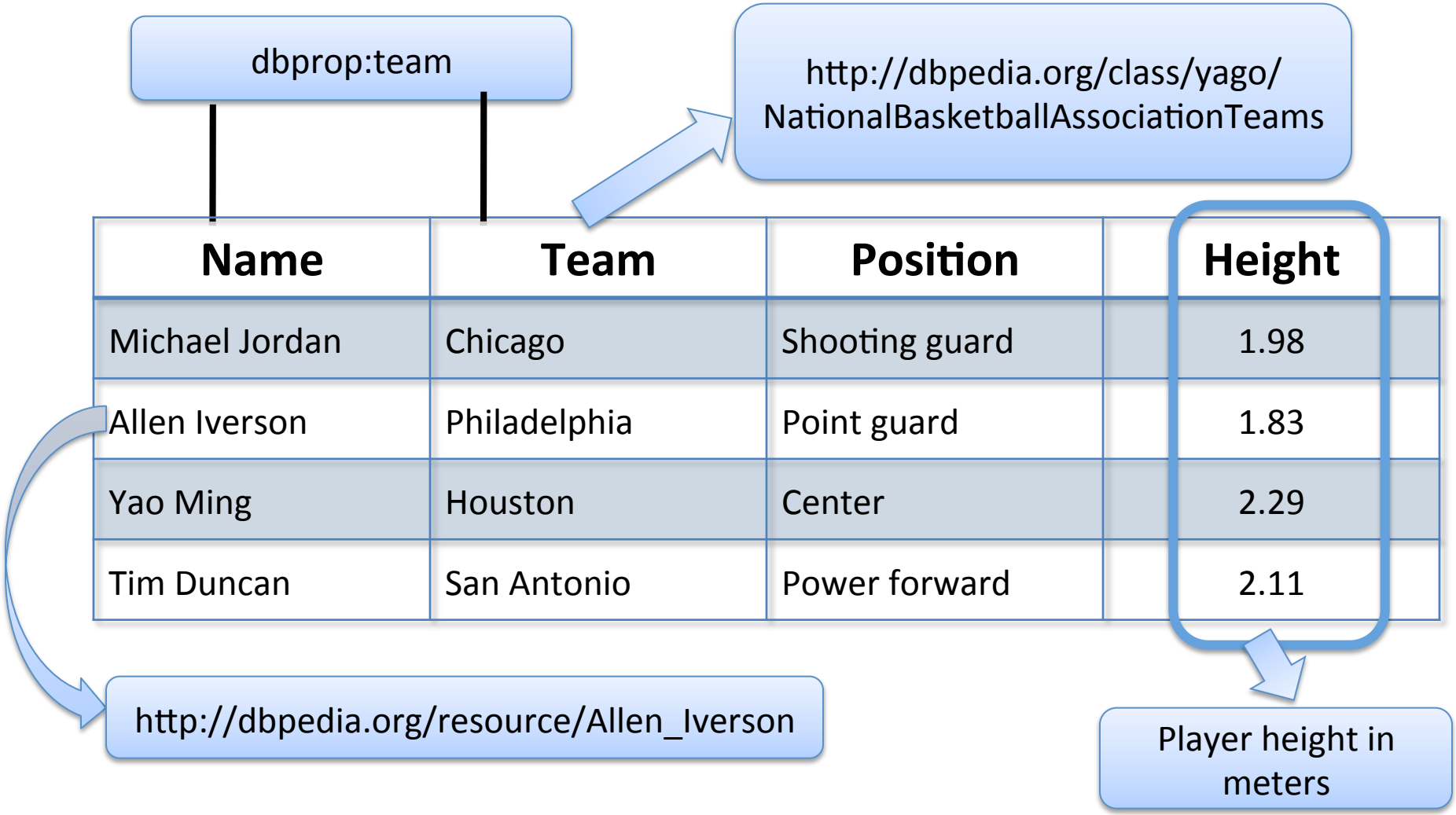
- Linked Open Data is hard to create
- Linded Open Data is hard to query
- Two ongoing UMBC dissertations hope to make it easier
  - Varish Mulwad: Generating linked data from tables
  - Lushan Han: Querying linked data with a quasi-NL interface
- Both need statistics on large amounts of LOD data and/or text

Generating Linked Data by Inferring the Semantics of Tables

Research by Varish Mulwad

http://ebiq.org/j/96

Goal: Table => LOD*



* DBpedia

Goal: Table => LOD*

Name	Team	Position	Height
Michael Jordan	Chicago	Shooting guard	1.98
Allen Iverson	Philadelphia	Point guard	1.83
Yao Ming	Houston		
Tim Duncan	San Anto		

RDF Linked Data

```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .
@prefix yago: <http://dbpedia.org/class/yago/> .

"Name"@en is rdfs:label of dbpedia-owl:BasketballPlayer .
"Team"@en is rdfs:label of yago:NationalBasketballAssociationTeams .

"Michael Jordan"@en is rdfs:label of dbpedia:Michael Jordan .
dbpedia:Michael Jordan a dbpedia-owl:BasketballPlayer .

"Chicago Bulls"@en is rdfs:label of dbpedia:Chicago Bulls .
dbpedia:Chicago Bulls a yago:NationalBasketballAssociationTeams .
```

All this in a completely automated way

* DBpedia

2010 Preliminary Heuristic System

T2LD Framework

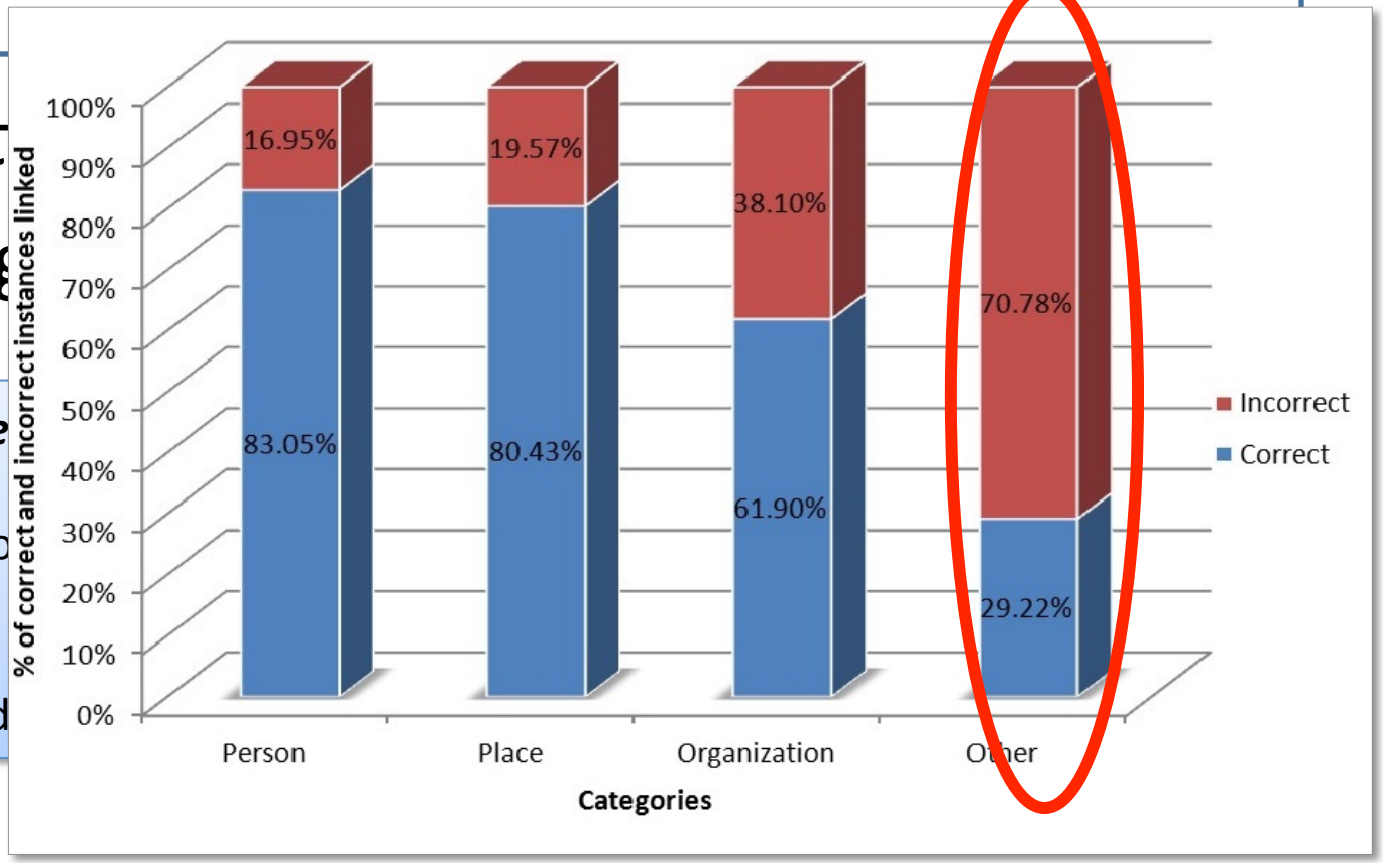
Predict Class for Columns → Linking the table cells → Identify and Discover relations

Class predict

Entity Linking

Examples of class labe
Column – Nationality
Prediction – MilitaryCo

Column – Birth Place
Prediction – Populated



Domain Independent Framework

City	Mayor	State	Population
Boston	T. Menino	MA	610,000
New York	M. Bloomberg	NY	8,400,000
Philadelphia	M. Nutter	PA	1,500,000
Baltimore	S. Dixon	MD	640,000
Washington	A. Fenty	DC	595,000

Query

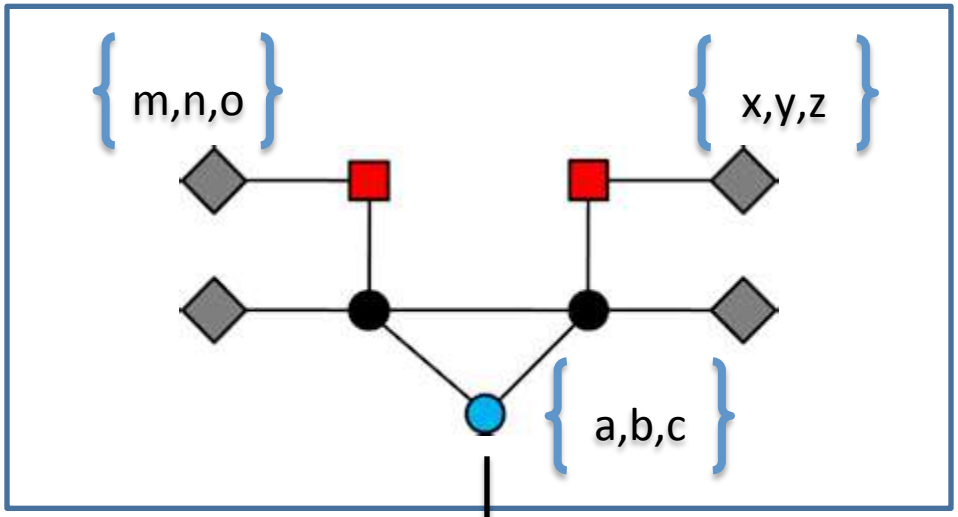
Generate a set of classes/entities

Domain Knowledge – Linked Data Cloud / Medical Domain / Open Govt. Domain

KB

KB

Wikitology

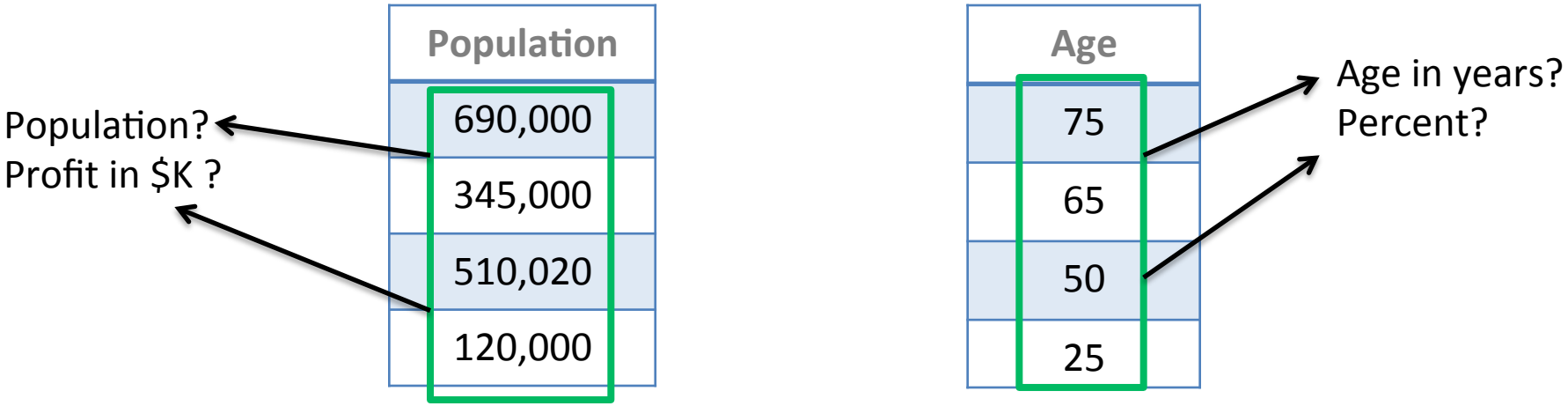


Joint Assignment/ Inference Model

Linked Data

One 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, e.g., age of *people* vs. *political leaders* vs. *professional athletes*
  - Represent as *measurements*: value + units
- Metric: possibility/probability of values given distribution

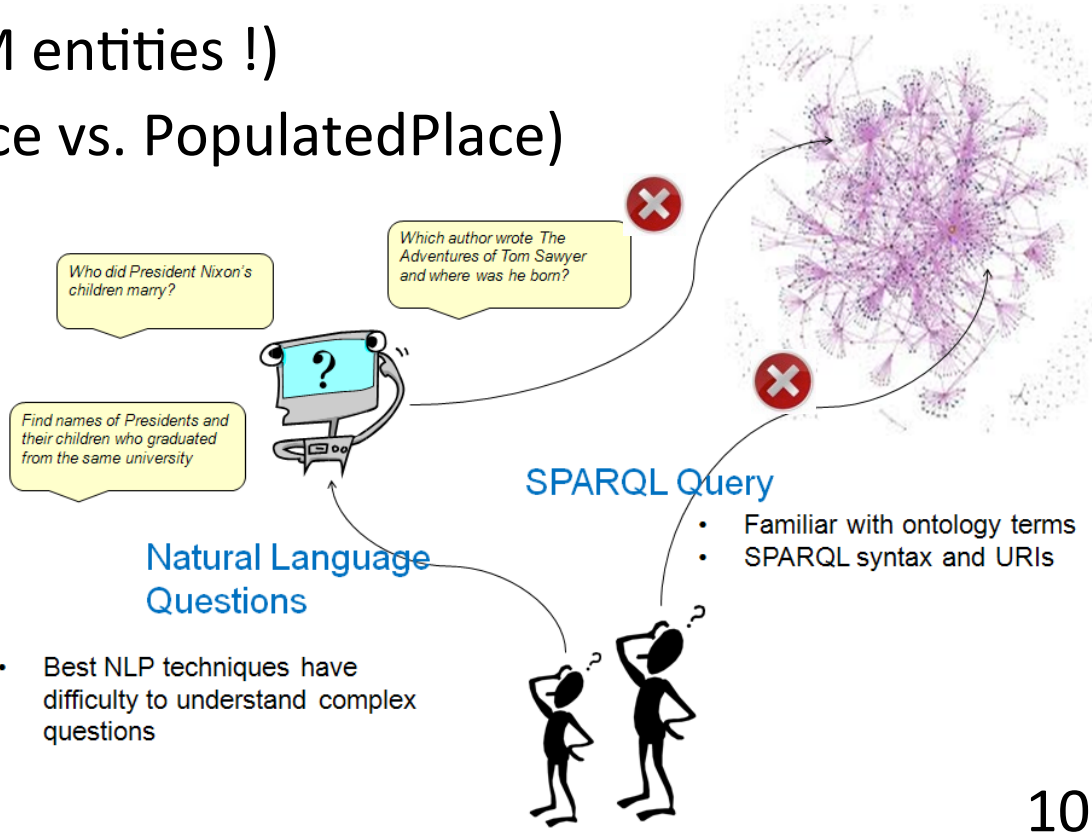
GoRelations:
Intuitive Query System for Linked Data

Research By Lushan Han

http://ebiq.org/j/93

Querying LOD is Too Hard

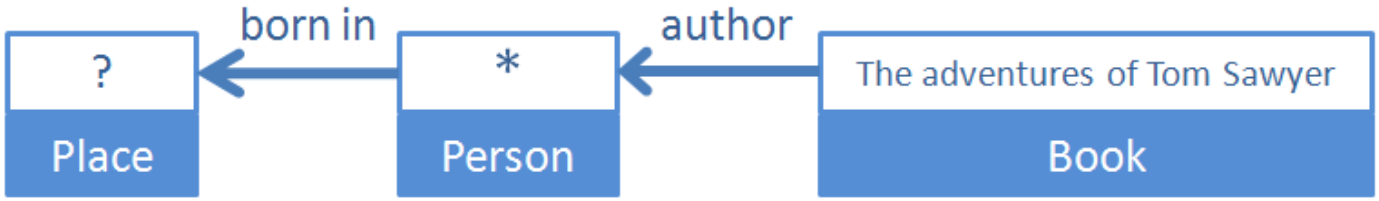
- Querying DBpedia requires a lot of a user
  - Understand the RDF model
  - Master SPARQL, a formal query language
  - Understand ontology terms: 320 classes & 1600 properties !
  - Know instance URIs (>1M entities !)
  - Term heterogeneity (Place vs. PopulatedPlace)
- Querying large LOD sets overwhelming
- Natural language query systems still a research goal



A Pragmatic Solution

- **Goal:** allow users with a basic RDF understanding to query LOD collections
  - To explore what data is available
  - To get answers to questions
  - To create SPARQL queries for reuse or adaptation
- **Desiderata:** Easy, Accurate, Fast
- **Key idea:** Reduce problem complexity by having (1) User enter a simple graph, and (2) Annotate it words and phrases

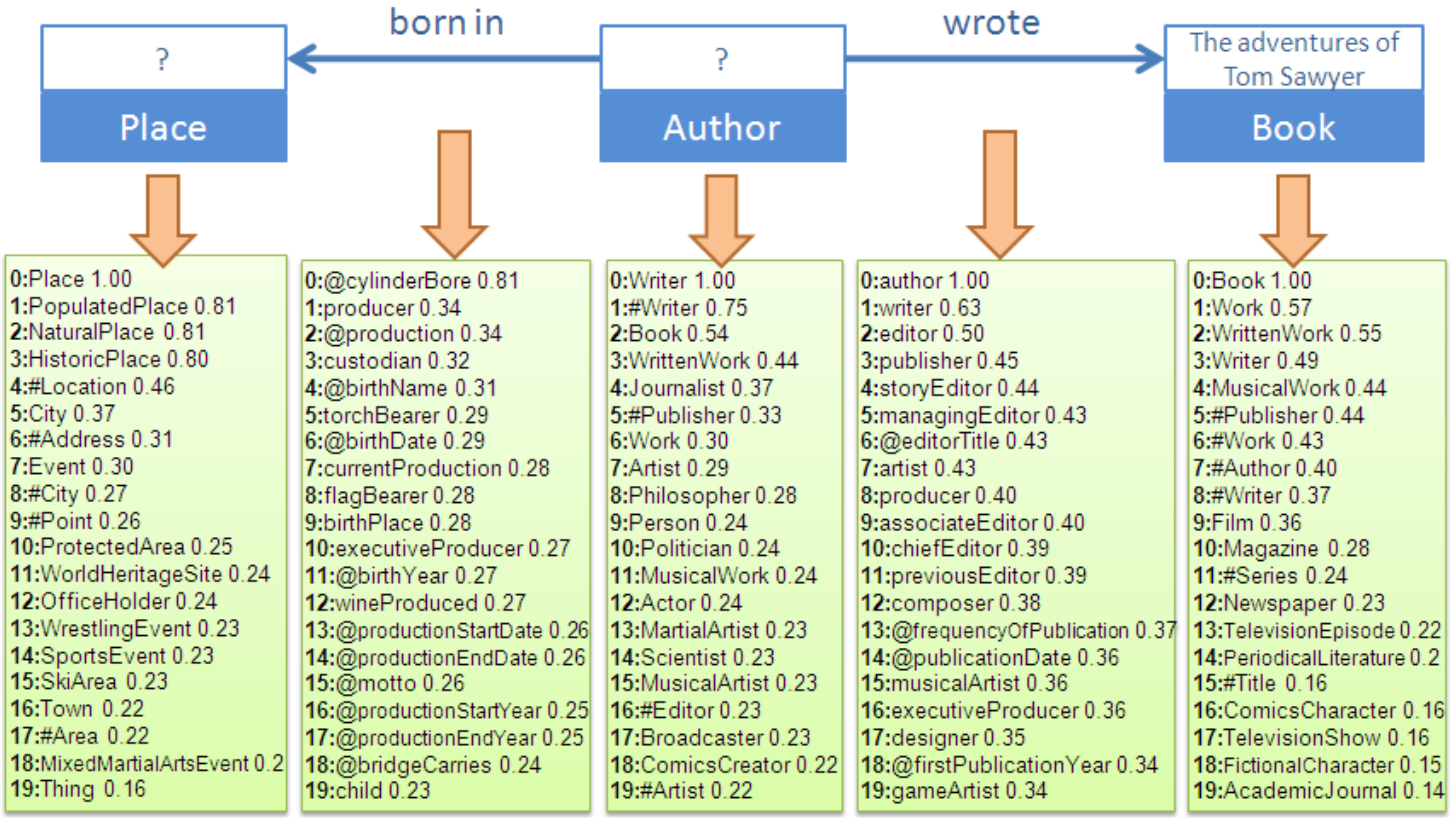
Semantic Graph Interface



- 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

Step 1: Find Terms

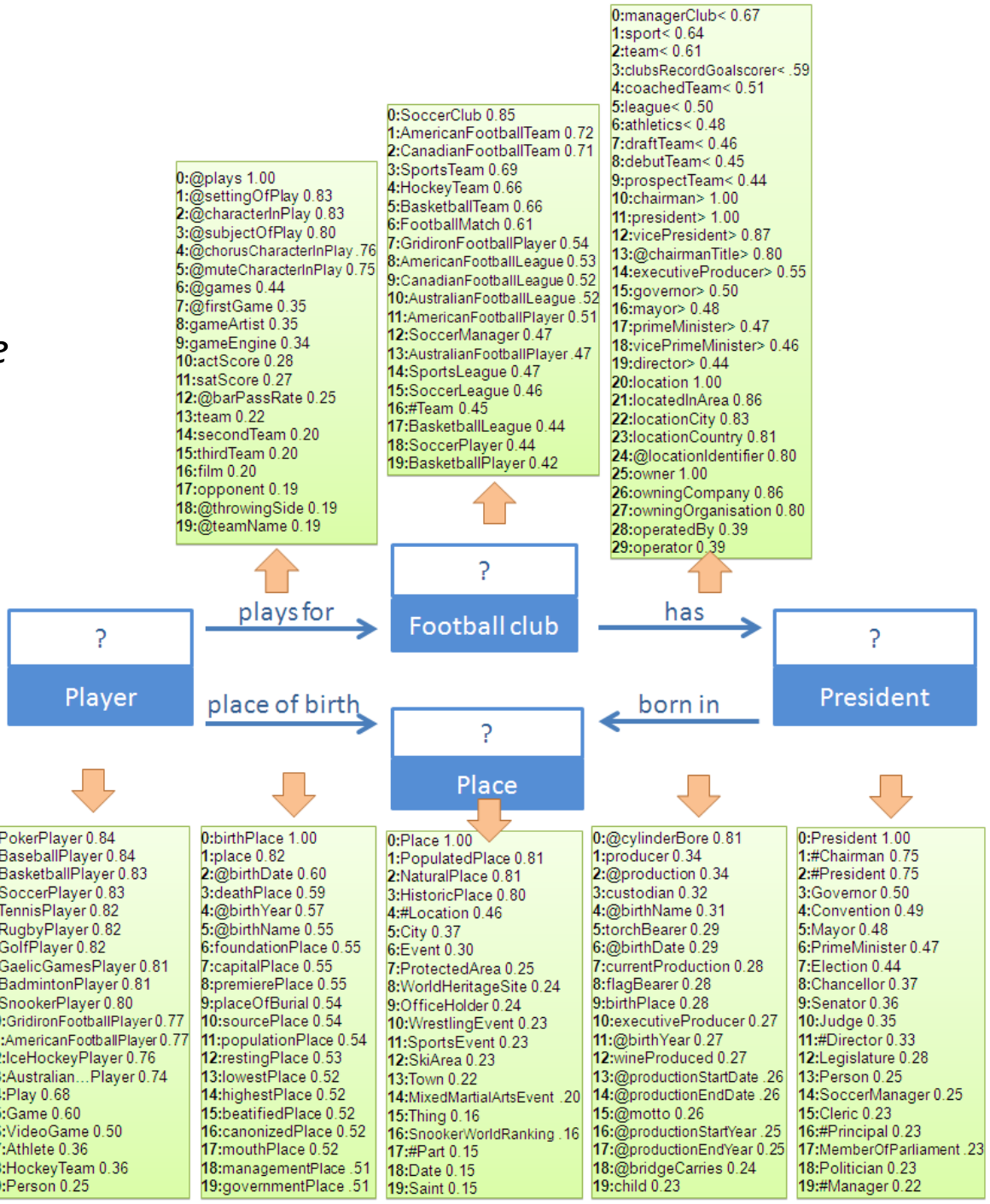
For each concept or relation in the graph, generate the *k* most semantically similar candidate ontology classes or properties



Similarity metric based on distributional similarity, LSA, and WordNet.

Another Example

Football players who were born in the same place as their team's president



Step 2: Disambiguate

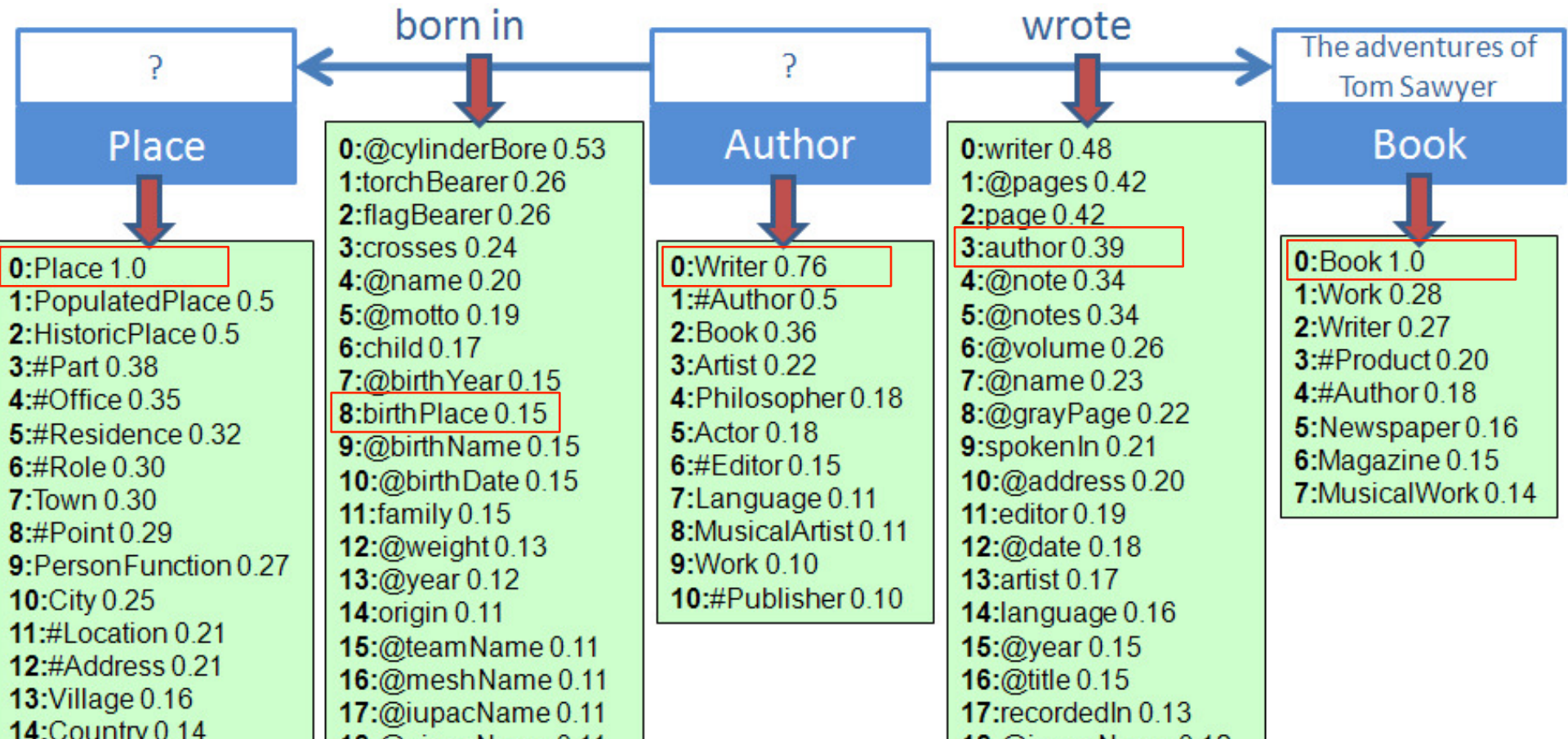
- Assemble the best interpretation using *statistics of the RDF data*
- Primary measure is *pointwise mutual information* (PMI) between RDF terms in LOD collection

  This measures the degree to which two RDF terms or types occur together in the knowledge base
- In good interpretations, *ontology terms associate* in the way that their corresponding *user terms* connect in the semantic graph

Example of Translation result

Concepts: Place => Place, Author => Writer, Book => Book

Properties: born in => birthPlace, wrote => author (inverse direction)

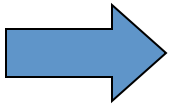


Step3: SPARQL Generation

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



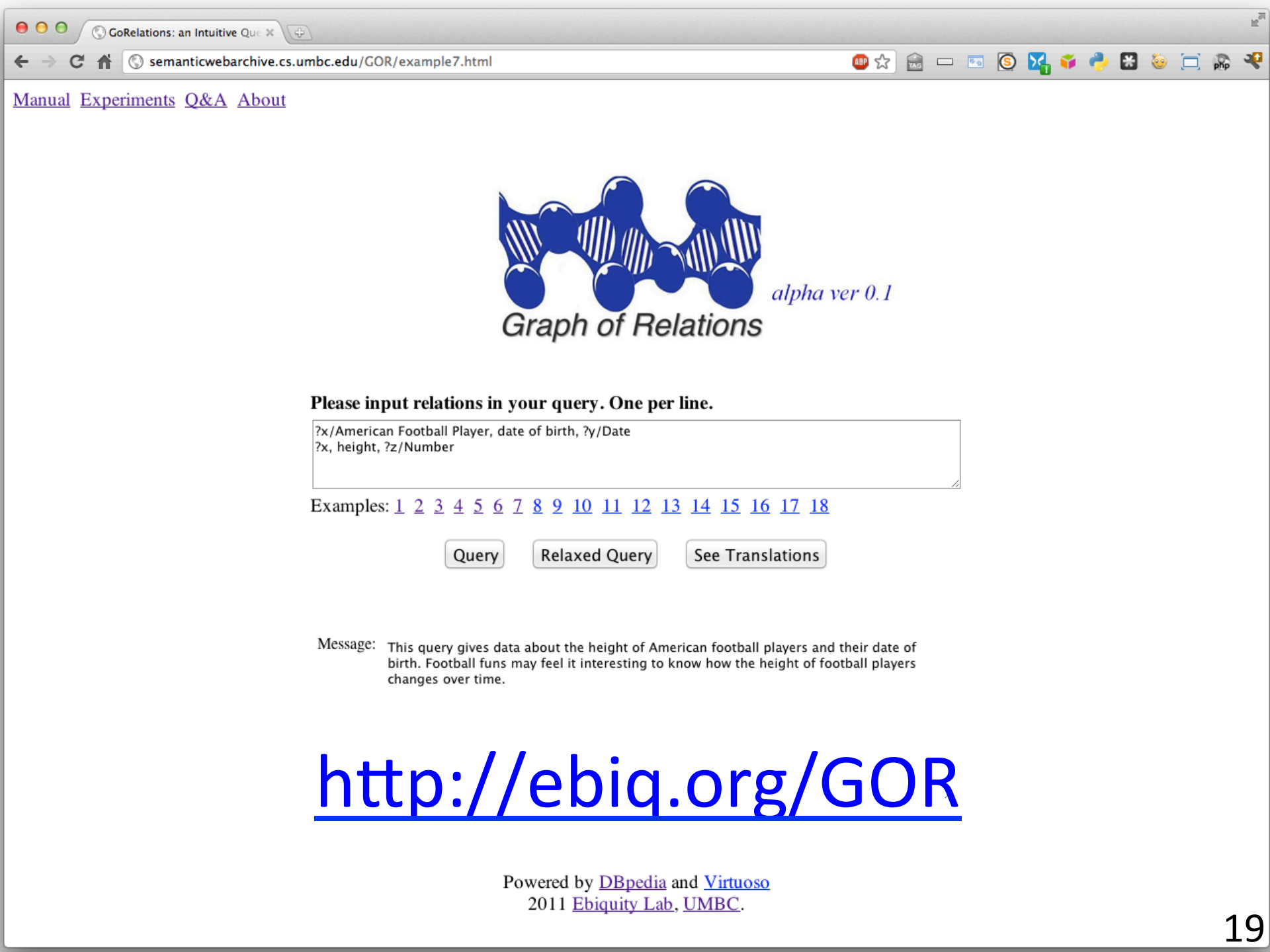
```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x, ?y WHERE {
  ?0 a dbo:Book .
  ?0 rdfs:label ?label0 .
  ?label0 bif:contains '"The adventures of Tom Sawyer"' .
  ?x a dbo:Writer .
  ?y a dbo:Place .
  {?0 dbo:author ?x} .
  {?x dbo:birthPlace ?y} .
}
```

Preliminary 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

		Prec.	*Recall*	*F*
GoRelations	regular	0.687	0.722	0.704
	concise	0.736	0.803	0.768
PowerAqua	1st triple	0.372	0.483	0.420
	all triples	0.334	0.483	0.395
	merged	0.255	0.291	0.272
True Knowledge		0.469	0.535	0.500


GoRelations: an Intuitive Que
semanticwebarchive.cs.umbc.edu/GOR/example7.html
Manual Experiments Q&A About
Graph of Relations
alpha ver 0.1
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 funs 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.

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