Making the Semantic Web Easier to Use

Tim Finin

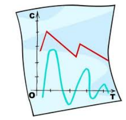
University of Maryland, Baltimore County

Joint work with Lushan Han, Varish Mulwad, Anupam Joshi

Presented at the Workshop on Data Engineering Meets the Semantic Web
of the IEEE International Conference on Data Engineering, 1 April 2012

UMBC ebiquity http://ebiq.org/r/339

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

2/49

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 2nd 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

3/49

Semantic Web

Use Semantic Web Technology to publish shared data & knowledge

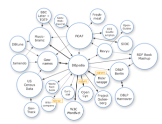
Semantic web technologies allow machines to share data and knowledge using common web language and protocols.

~ 1997

Semantic Web beginning

Semantic Web => Linked Open Data

Use Semantic Web Technology
to publish shared data &
knowledge



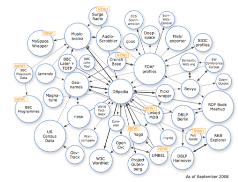
2007

Data is inter-
linked to support inte-
gration and fusion of knowledge

LOD beginning

Semantic Web => Linked Open Data

Use Semantic Web Technology
to publish shared data &
knowledge



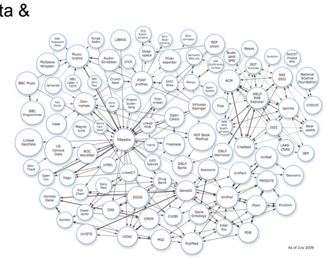
2008

Data is inter-
linked to support inte-
gration and fusion of knowledge

LOD growing

Semantic Web => Linked Open Data

Use Semantic Web Technology
to publish shared data &
knowledge



2009

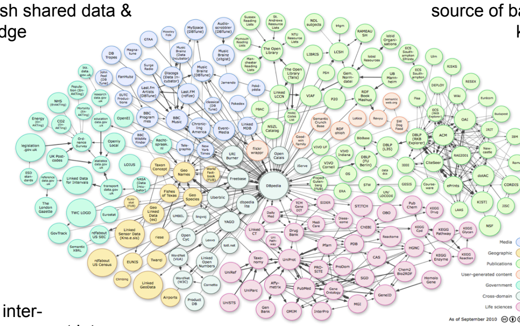
Data is inter-
linked to support inte-
gration and fusion of knowledge

… and growing

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 inter-
linked to support inte-
gration and fusion of knowledge

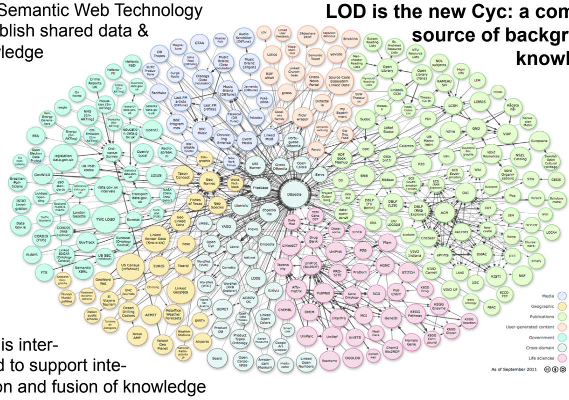
2010

…growing faster

Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

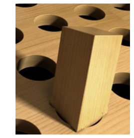
LOD is the new Cyc: a common source of background knowledge



Data is inter-linked to support inte-gration 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 appropri-ate 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

10/49

Generating Linked Data by Inferring the Semantics of Tables

Research with Varish Mulwad

http://ebiq.org/j/96

Goal: Table => LOD*

dbprop:team

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

Name	Team	Position	Height
Michael Jordan	Chicago	Shooting guard	1.98
Allen Iverson	Philadelphia	Point guard	1.83
Yao Ming	Houston	Center	2.29
Tim Duncan	San Antonio	Power forward	2.11

http://dbpedia.org/resource/Allen_Iverson

Player height in meters

* DBpedia

12/49

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 Ant[o]		

RDF Linked Data

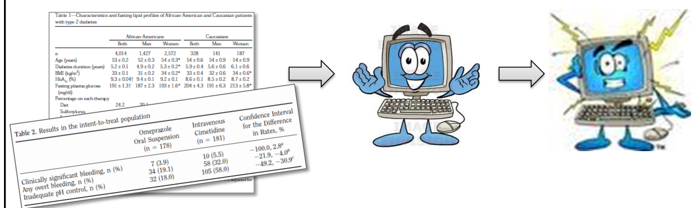
```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix yago: <http://dbpedia.org/class/yago/> .

"Name"@en is rdfs:label of dbo:BasketballPlayer .
"Team"@en is rdfs:label of yago:NationalBasketballAssociationTeams .

"Michael Jordan"@en is rdfs:label of dbpedia:Michael Jordan .
dbpedia:Michael Jordan a dbo: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

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

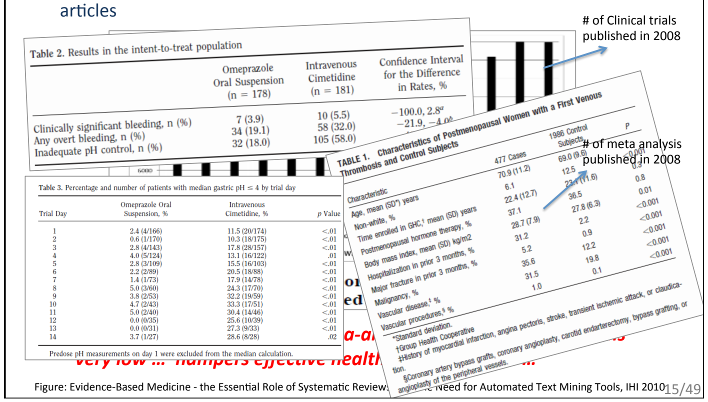


Table 2. Results in the intent-to-treat population

	Omeprazole Oral Suspension (n = 178)	Intravenous Cimetidine (n = 181)	Confidence Interval for the Difference in Rates, %
Clinically significant bleeding, n (%)	7 (3.9)	10 (5.5)	−100.0, 2.8[a]
Any overt bleeding, n (%)	34 (19.1)	58 (32.0)	−21.9, −4.0[b]
Inadequate pH control, n (%)	32 (18.0)	105 (58.0)	

Table 3. Percentage and number of patients with median gastric pH ≤ 4 by trial day

Trial Day	Omeprazole Oral Suspension, %	Intravenous Cimetidine, %	p Value
1	2.4 (4/166)	11.5 (20/174)	<.01
2	0.6 (1/170)	10.3 (18/175)	<.01
3	2.8 (4/143)	17.8 (28/157)	<.01
4	4.0 (5/124)	13.1 (16/122)	.01
5	2.8 (3/109)	15.5 (16/103)	<.01
6	2.2 (2/89)	20.5 (18/88)	<.01
7	1.4 (1/73)	17.9 (14/78)	<.01
8	5.0 (3/60)	24.3 (17/70)	<.01
9	3.8 (2/53)	32.2 (19/59)	<.01
10	4.7 (2/43)	33.3 (17/51)	<.01
11	5.0 (2/40)	30.4 (14/46)	<.01
12	0.0 (0/35)	25.6 (10/39)	<.01
13	0.0 (0/31)	27.3 (9/33)	<.01
14	3.7 (1/27)	28.6 (8/28)	.02

Predose pH measurements on day 1 were excluded from the median calculation.

TABLE 1. Characteristics of Postmenopausal Women with a First Venous Thrombosis and Control Subjects

Figure: Evidence-Based Medicine - the Essential Role of Systematic Review[s] … [Th]e Need for Automated Text Mining Tools, IHI 2010

DATA.GOV
EMPOWERING PEOPLE

~ ***400,000*** datasets

~ ***< 1*** % in RDF

2010 Preliminary System

T2LD Framework

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

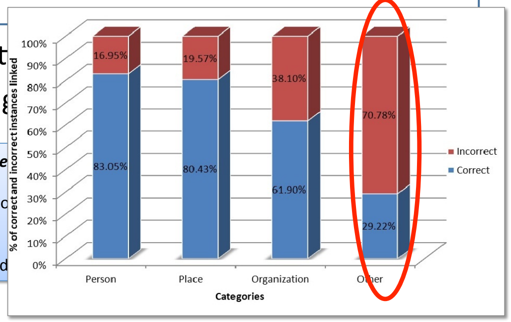
Class predict

Entity Linkin

Examples of class labe

Column – Nationality
Prediction – MilitaryCo

Column – Birth Place
Prediction – Populated

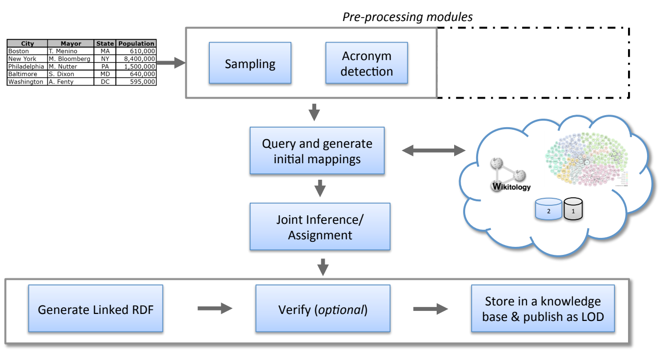


Sources of Errors

- 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

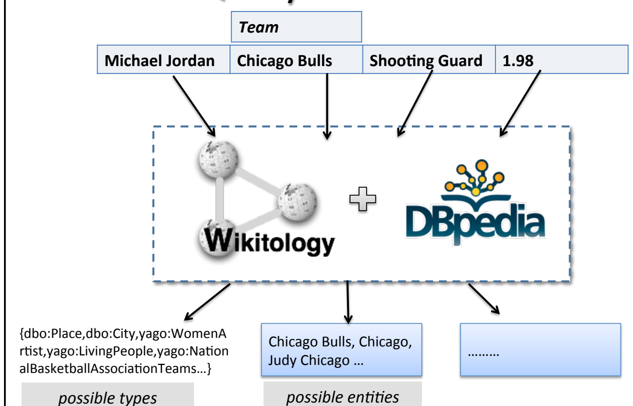
18/49

A Domain Independent Framework



19/49

Query Mechanism



20/49

Ranking the candidates

- C_i = "State" ; L_{ci} = AdministrativeRegion

String in column header

Class from an ontology

- f_1 = [Levenshtein distance(C_i , L_{Ci}),
  Dice Score (C_i , L_{Ci}),
  Semantic Similarity (C_i , L_{Ci}),
  InformationGain(L_{Ci})]

String similarity metrics

- $\psi_1 = \exp(w_1^T f_1(C_i , L_{Ci}))$

Ranking the candidates

- R_{ij} = "Baltimore" ; E_{ij} = Baltimore_Maryland

String in table cell

Entity from the knowledge base (KB)

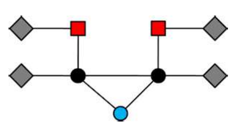
- f_2 = [Levenshtein distance(R_{ij} , E_{ij}),
  Dice Score (R_{ij} , E_{ij}),
  PageRank (E_{ij}),
  KBScore (E_{ij})
  PageLength (E_{ij})]

String similarity metrics

Popularity metrics

$\psi_2 = \exp(w_2^T f_2(R_{ij} , E_{ij}))$

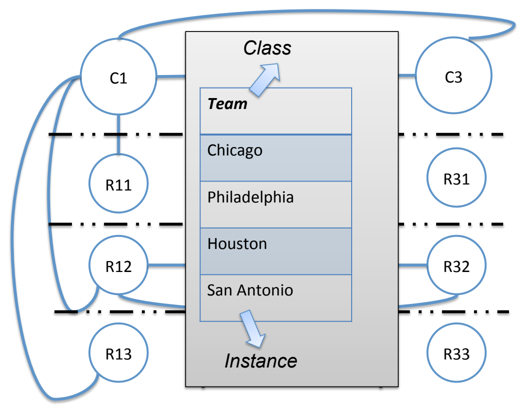
Joint Inference over evidence in a table



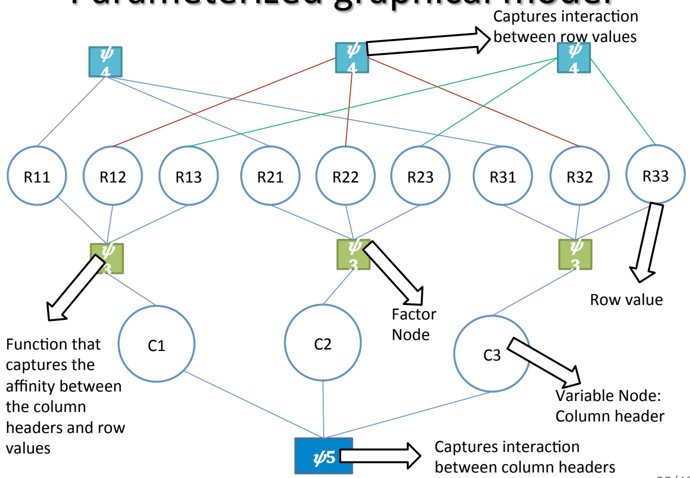
✓Probabilistic Graphical Models

A graphical model for tables

Joint inference over evidence in a table



Parameterized graphical model



Challenge: Interpreting Literals

Many columns have literals, e.g., numbers

Population
690,000
345,000
510,020
120,000

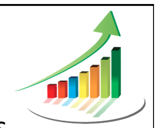
Population? Profit in $K ?

Age
75
65
50
25

Age in years? Percent?

- 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

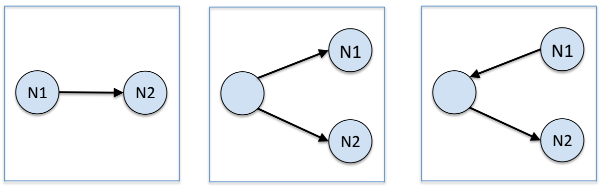
We use pointwise mutual information (pmi) to measure the association between two RDF resources (nodes)

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

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

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

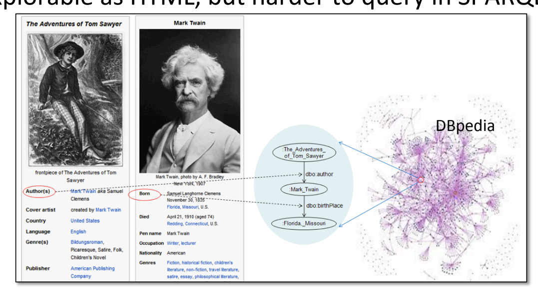
GoRelations:
Intuitive Query System for Linked Data

Research with Lushan Han

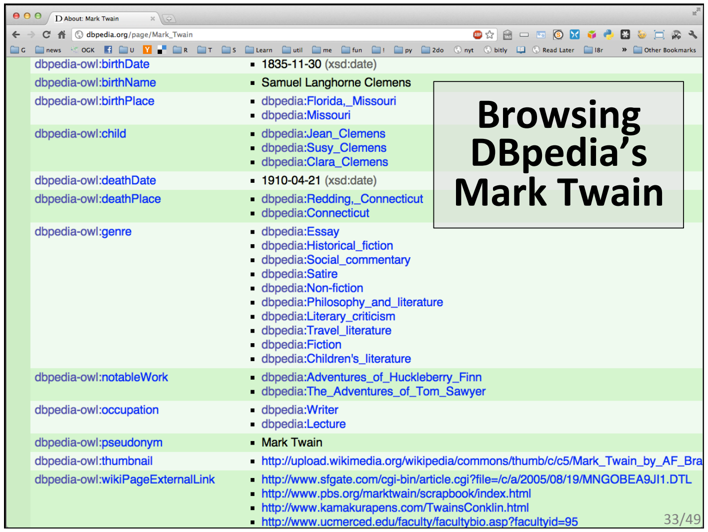
http://ebiq.org/j/93

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

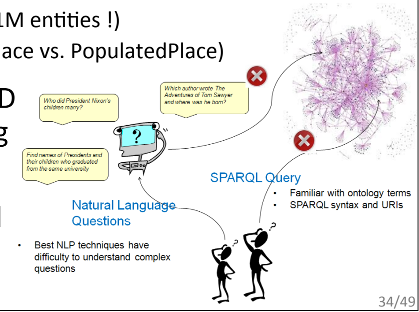


Browsing DBpedia's Mark Twain



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 distrib-uted 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

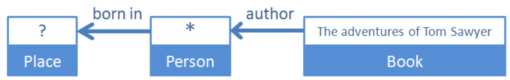
Key Idea

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

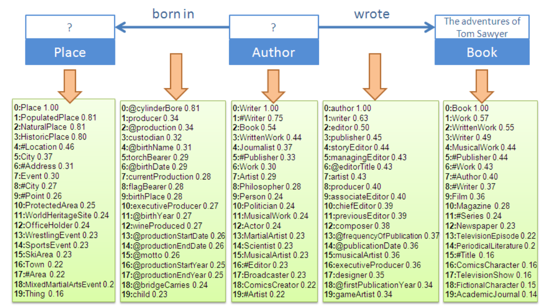


- Nodes denote entities and links binary relations
- Entities described by two unrestricted terms: *name* or value and *type* or concept
- Result entities marked with *?* and those not with *
- A compromise between a natural language Q&A system and SPARQL
  - Users provide compositional structure of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

Translation – Step One

finding semantically similar ontology terms

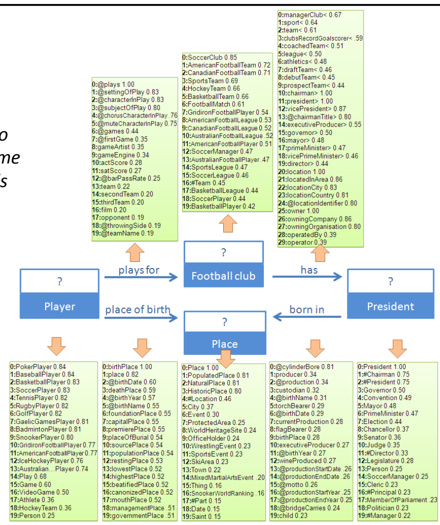
For each concept or relation in the graph, generate the *k* most semantically similar candidate ontology classes or properties



Similarity metric is distributional similarity, LSA, and WordNet.

Another Example

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



Translation – Step Two

disambiguation algorithm

- 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

$$\underset{p_1..p_m\; c_1..c_n \in H}{\operatorname{argmax}} \text{goodness}(G) = \underset{p_1..p_m\; c_1..c_n \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \text{goodness}(L_i) \quad (1)$$

Resolving direction

$$\begin{aligned} &\text{If } [\overrightarrow{\mathrm{PMI}}(c(O_i), p(R_i)) + \overrightarrow{\mathrm{PMI}}(p(R_i), c(S_i))] \\ &\quad - [\overrightarrow{\mathrm{PMI}}(c(S_i), p(R_i)) + \overrightarrow{\mathrm{PMI}}(p(R_i), c(O_i))] > \alpha \\ &\text{Then } S_i' = O_i,\ O_i' = S_i \\ &\text{Else } S_i' = S_i,\ O_i' = O_i \end{aligned} \quad (2)$$

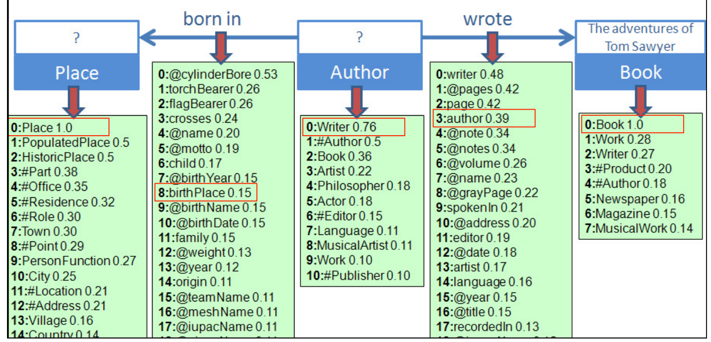
Link reasonableness

$$\begin{aligned} \text{goodness}(L_i) = \max(&\overrightarrow{\mathrm{PMI}}(c(S_i'), p(R_i)) \cdot \text{sim}(S_i', c(S_i')) \cdot \text{sim}(R_i, p(R_i)) \\ &+ \overrightarrow{\mathrm{PMI}}(p(R_i), c(O_i')) \cdot \text{sim}(O_i', c(O_i')) \cdot \text{sim}(R_i, p(R_i)), \beta) \\ &+ \mathrm{PMI}(c(S_i'), c(O_i')) \cdot \text{sim}(S_i', c(S_i')) \cdot \text{sim}(O_i', c(O_i')) \end{aligned} \quad (3)$$

Example of Translation result

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

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



SPARQL Generation

The translation of a semantic graph query to SPARQL is straightforward given the mappings

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

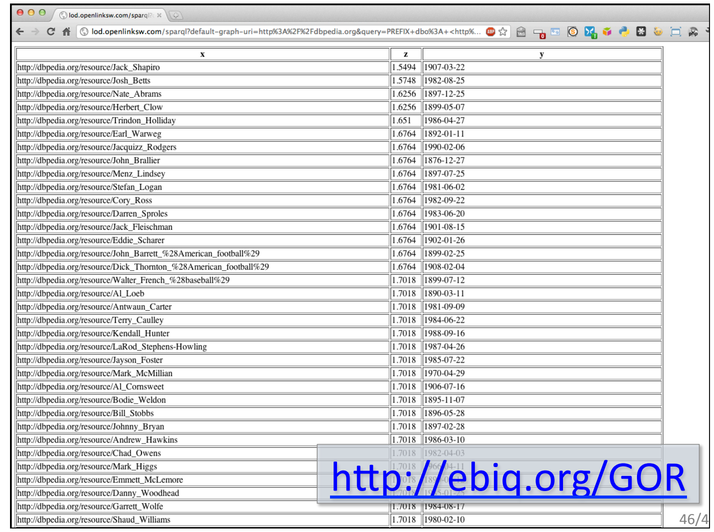
Relations
- born in => birthPlace
- wrote => author

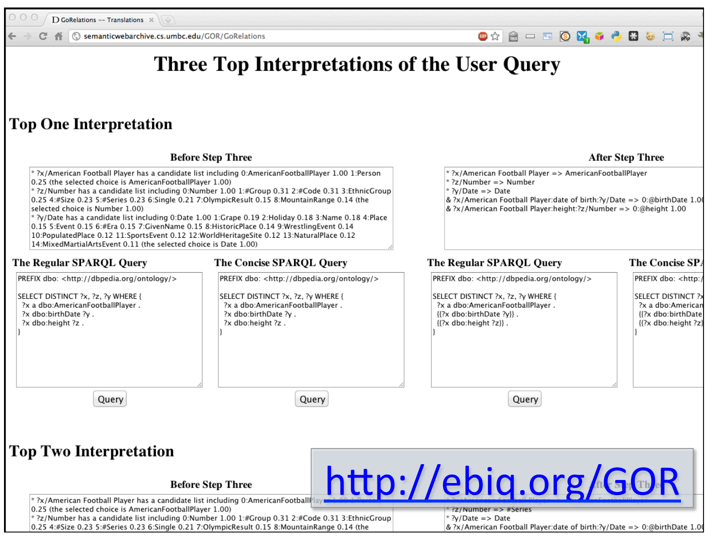
```
PREFIX dbo: <http://dbpedia.org/ontology/>

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} .
}
```


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





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

48/49

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

49/49

http://ebiq.org/