A Multidimensional Semantic Space for Data Model Independent Queries over RDF Data

André Freitas, João Gabriel Oliveira, Edward Curry
Seán O'Riain

Outline

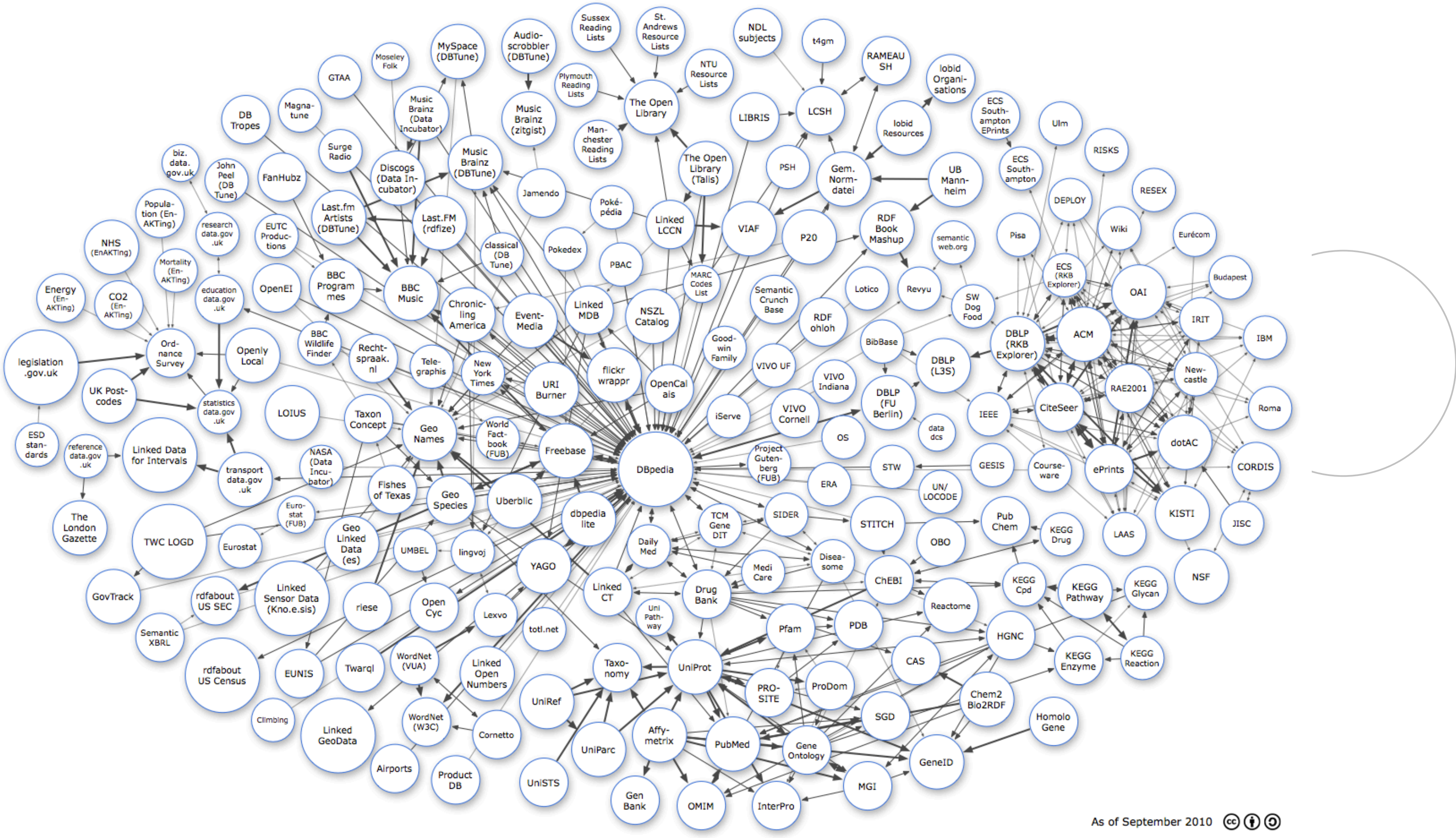
- Problem Space & Motivation
- Description of the Approach
- Evaluation
- Conclusion & Future Work

Linked Data

- Uses the Web infrastructure and standards to expose and interlink datasets.

- Linked Data vision:
  - □ The Web as a single Dataspace.
  - □ Web of interlinked datasets.

Linked Data: Adoption



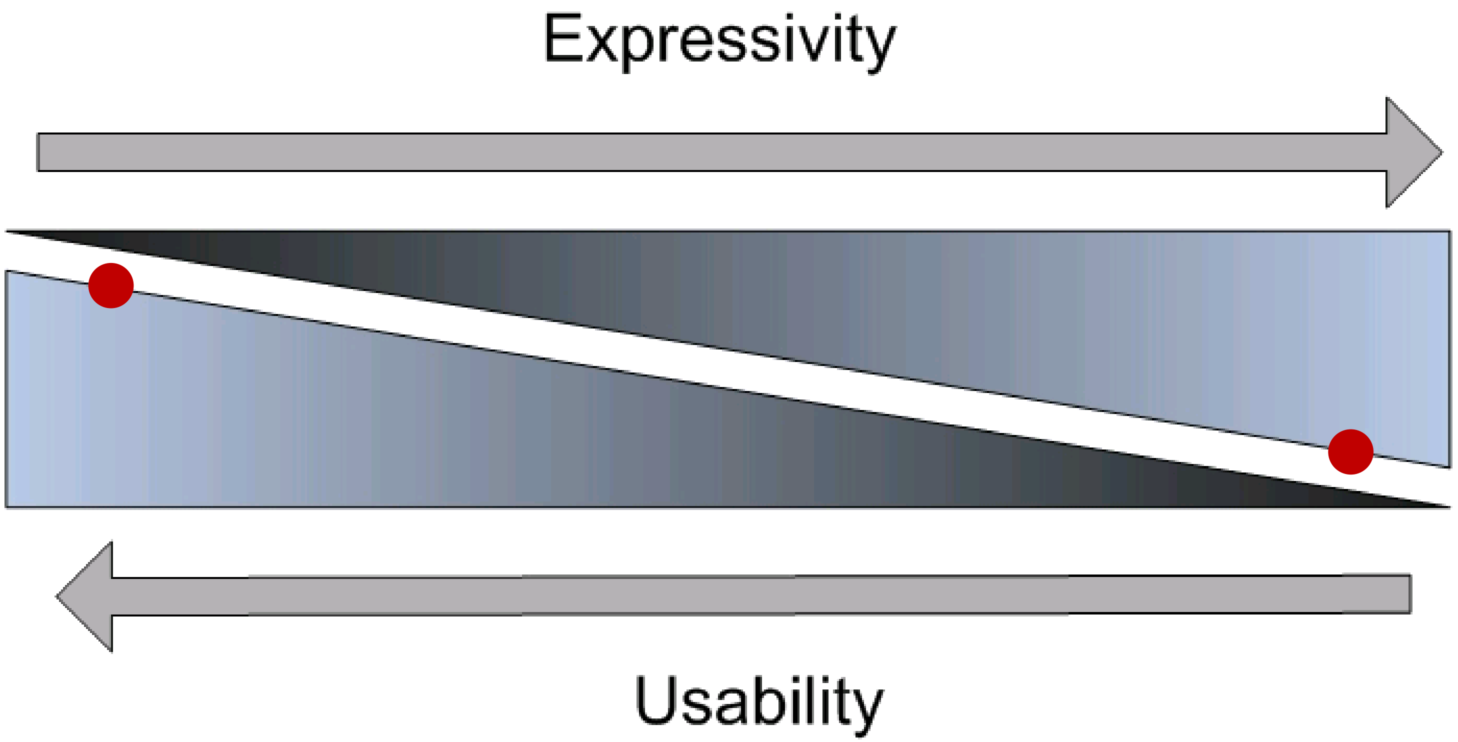
As of September 2010

Queries over Linked Data

- Linked Data brings a fundamental challenge for data consumption:
  - □ How to query heterogeneous and distributed datasets?
  - □ At Web scale it is unfeasible for end-users to be aware of the location and structure of datasets.

- Demand for new query mechanisms for Linked Data (data model independency).

Query/Search Spectrum



Expressivity

Usability

Keyword search

Entiy-centric search

...

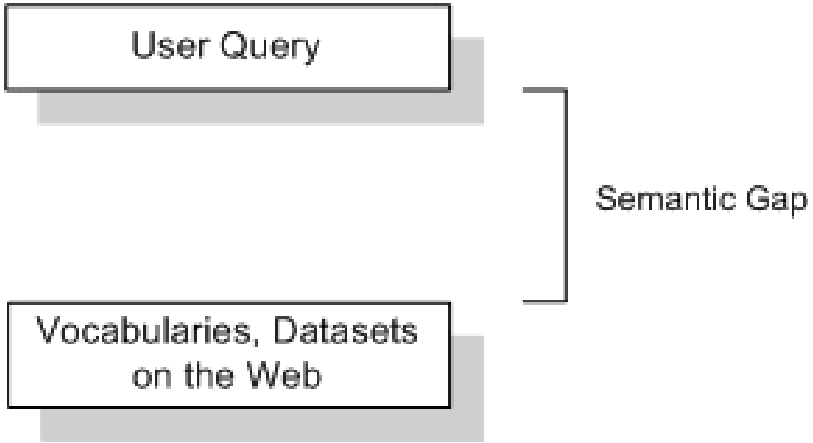
Visual query interfaces

SPARQL

Adapted from Kauffman et al (2009)

Fundamental Problem

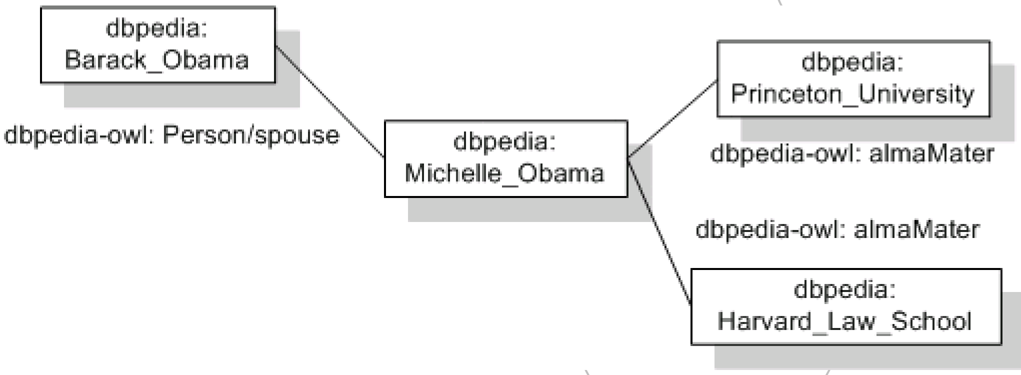
From which university did the wife of Barack Obama graduate?



User Query

Semantic Gap

Vocabularies, Datasets on the Web



dbpedia: Barack_Obama

dbpedia-owl: Person/spouse

dbpedia: Michelle_Obama

dbpedia-owl: almaMater

dbpedia: Princeton_University

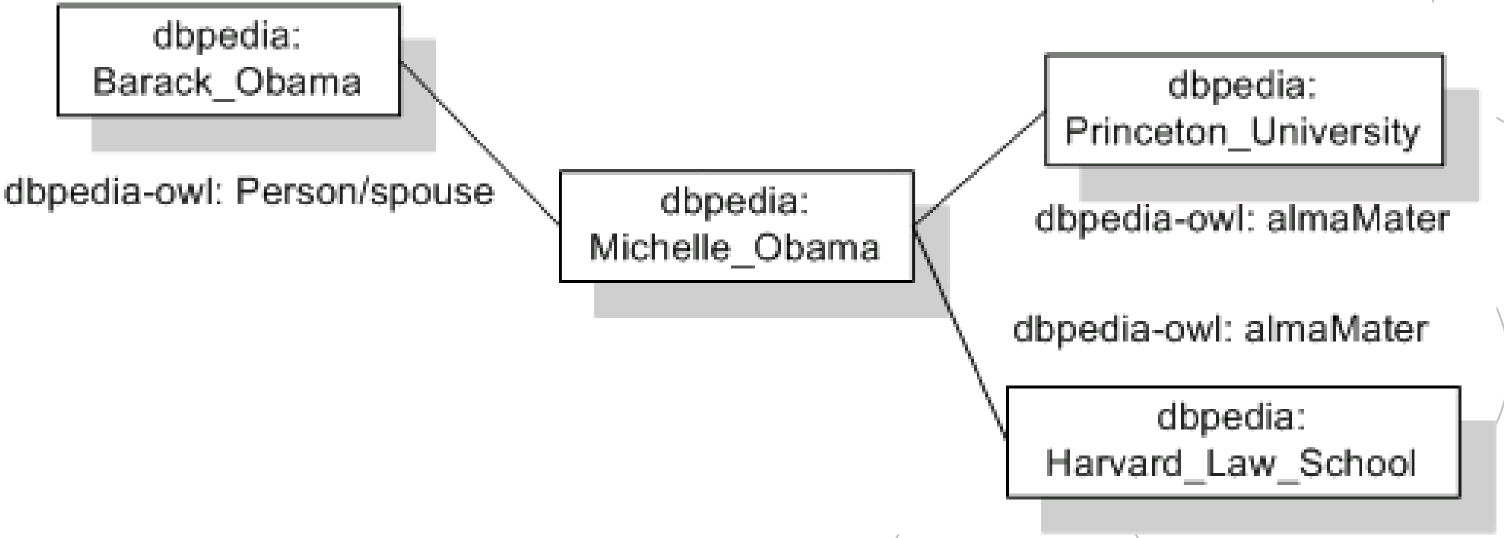
dbpedia-owl: almaMater

dbpedia: Harvard_Law_School

- Popescu (2003): Semantic tractability problem.

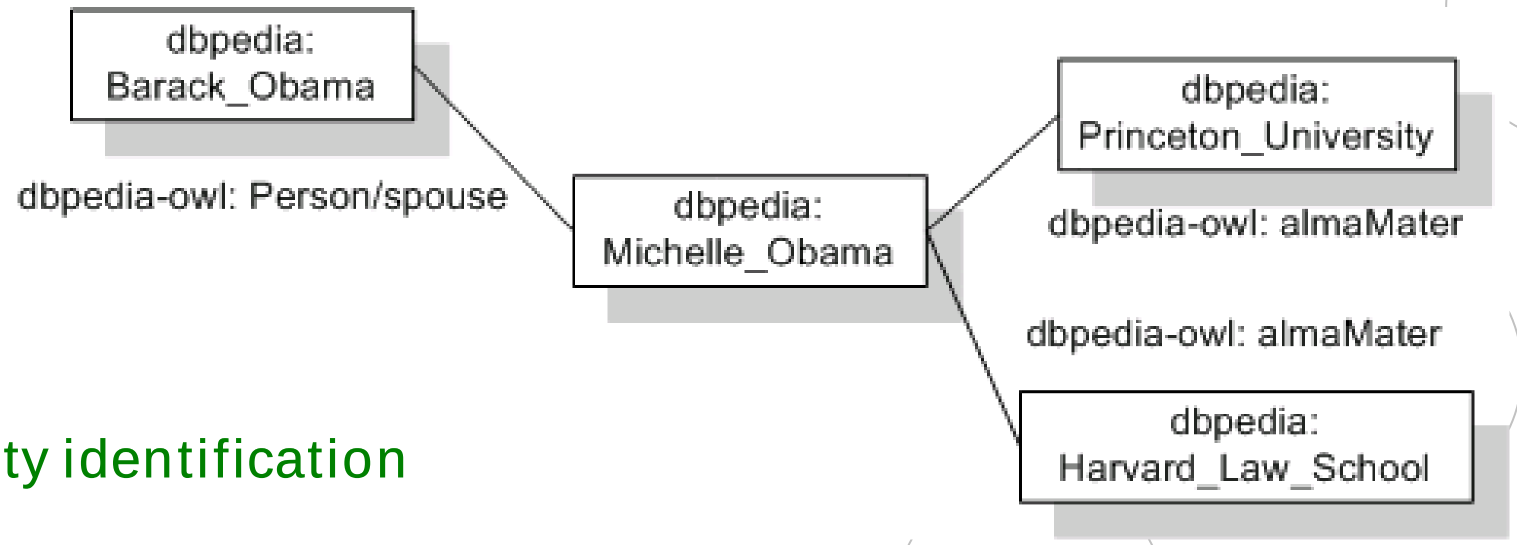
Semantic Matching Problem

From which university did the wife of Barack Obama graduate?



Semantic Matching Problem

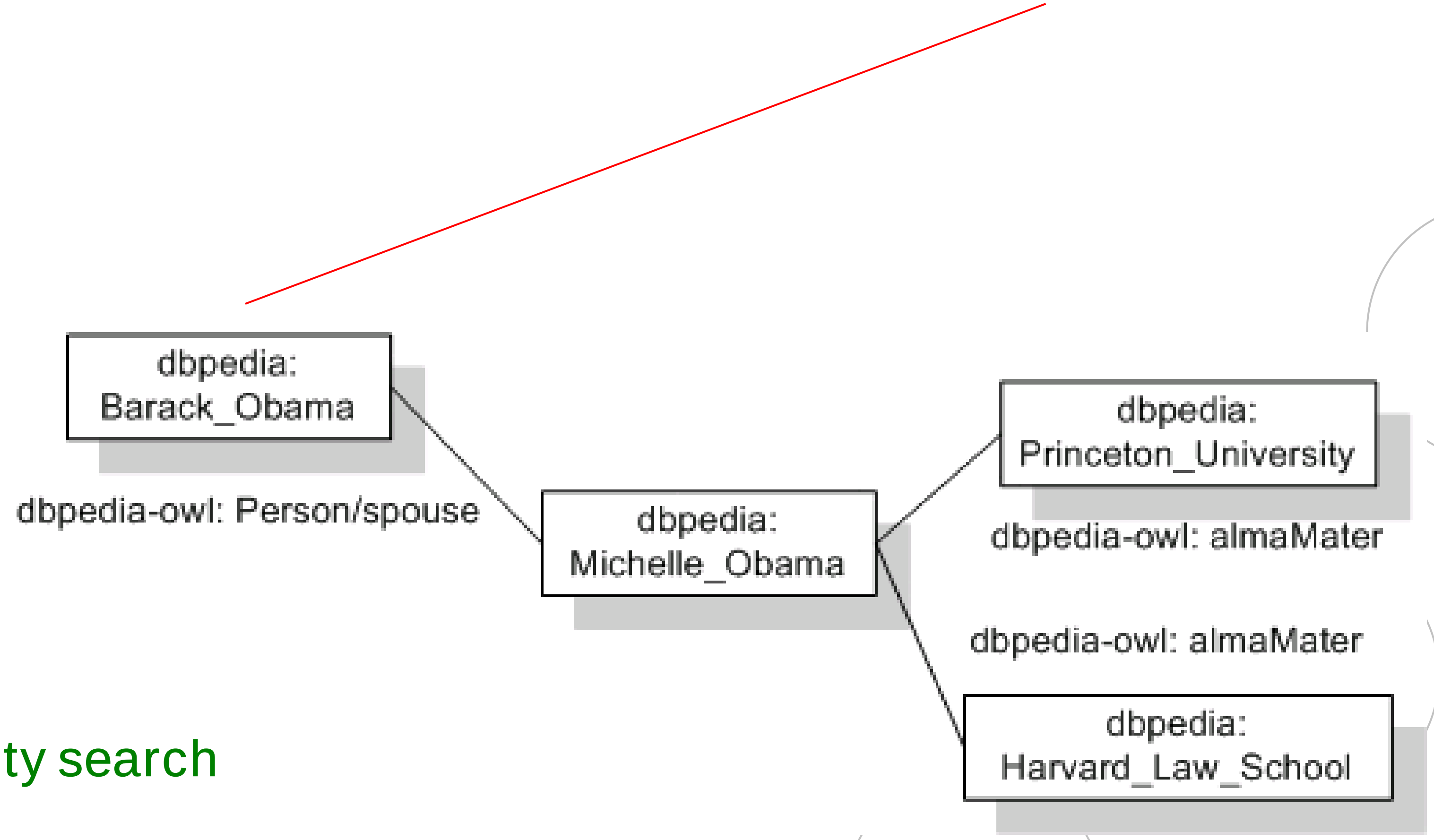
From which university did the wife of Barack Obama graduate?



Entity identification

Semantic Matching Problem

From which university did the wife of Barack Obama graduate?



dbpedia: Barack_Obama

dbpedia-owl: Person/spouse

dbpedia: Michelle_Obama

dbpedia: Princeton_University

dbpedia-owl: almaMater

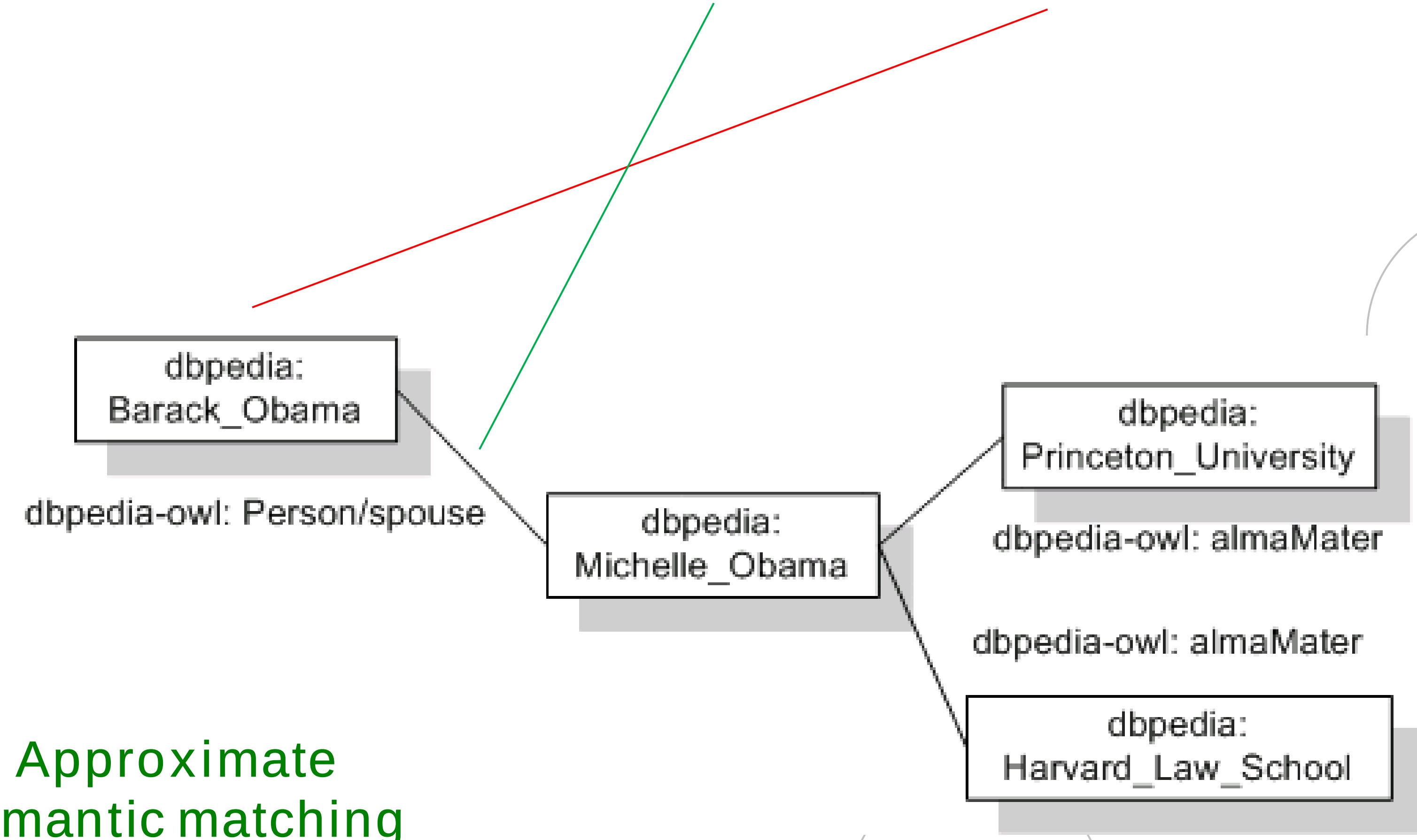
dbpedia-owl: almaMater

dbpedia: Harvard_Law_School

Entity search

Semantic Matching Problem

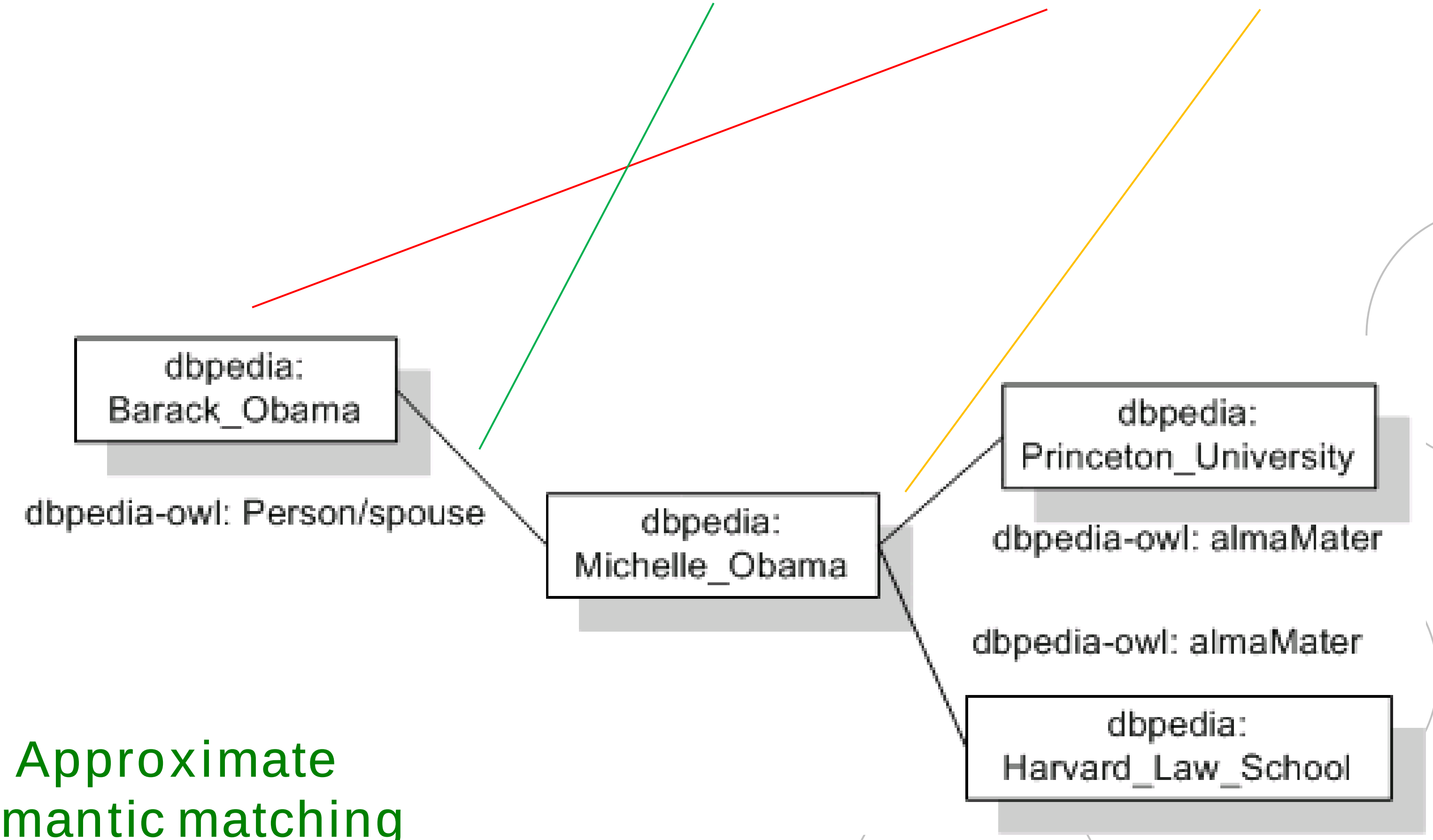
From which university did the wife of Barack Obama graduate?



Approximate semantic matching

Semantic Matching Problem

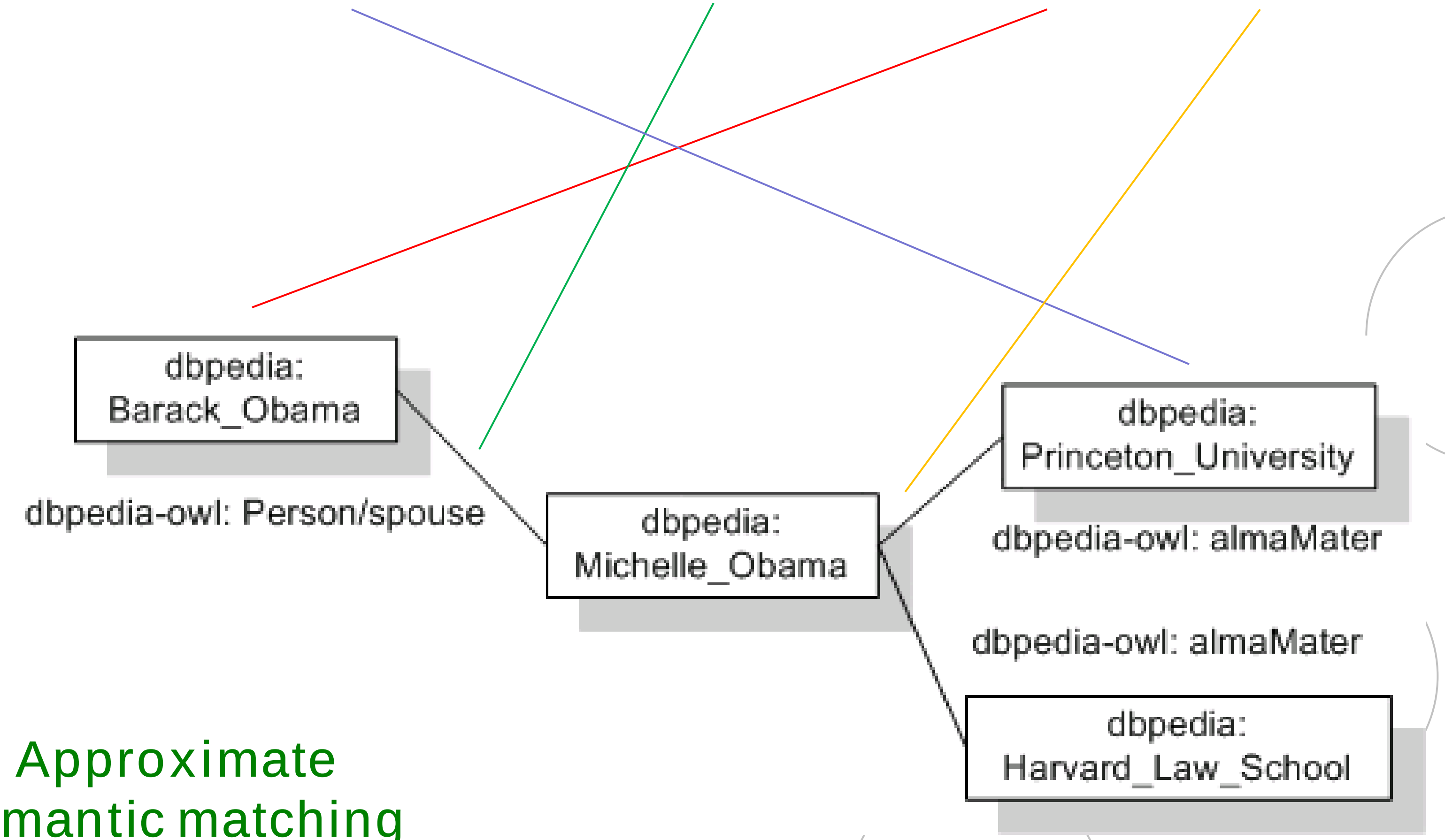
From which university did the wife of Barack Obama graduate?



Approximate semantic matching

Semantic Matching Problem

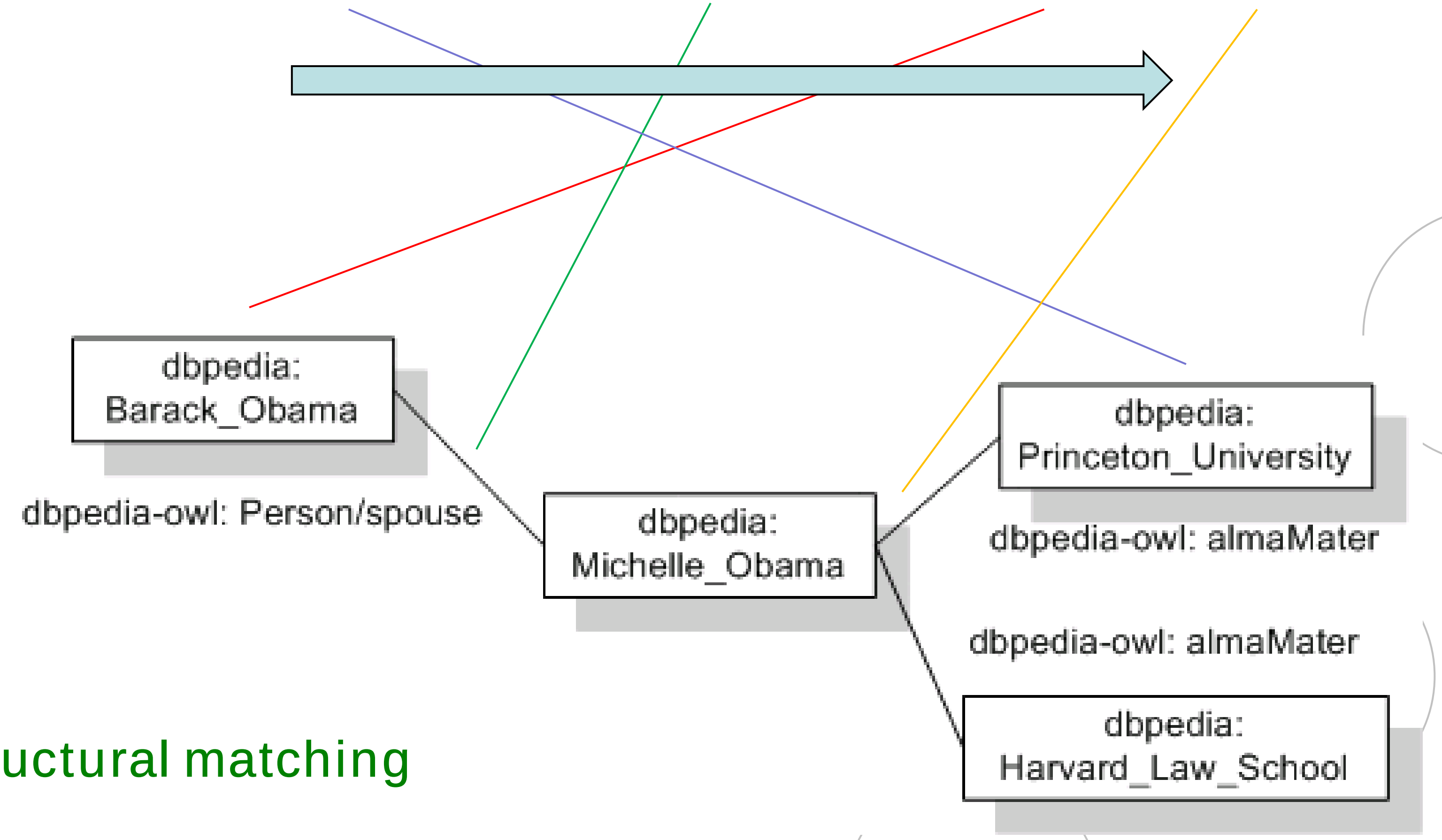
From which university did the wife of Barack Obama graduate?



Approximate semantic matching

Semantic Matching Problem

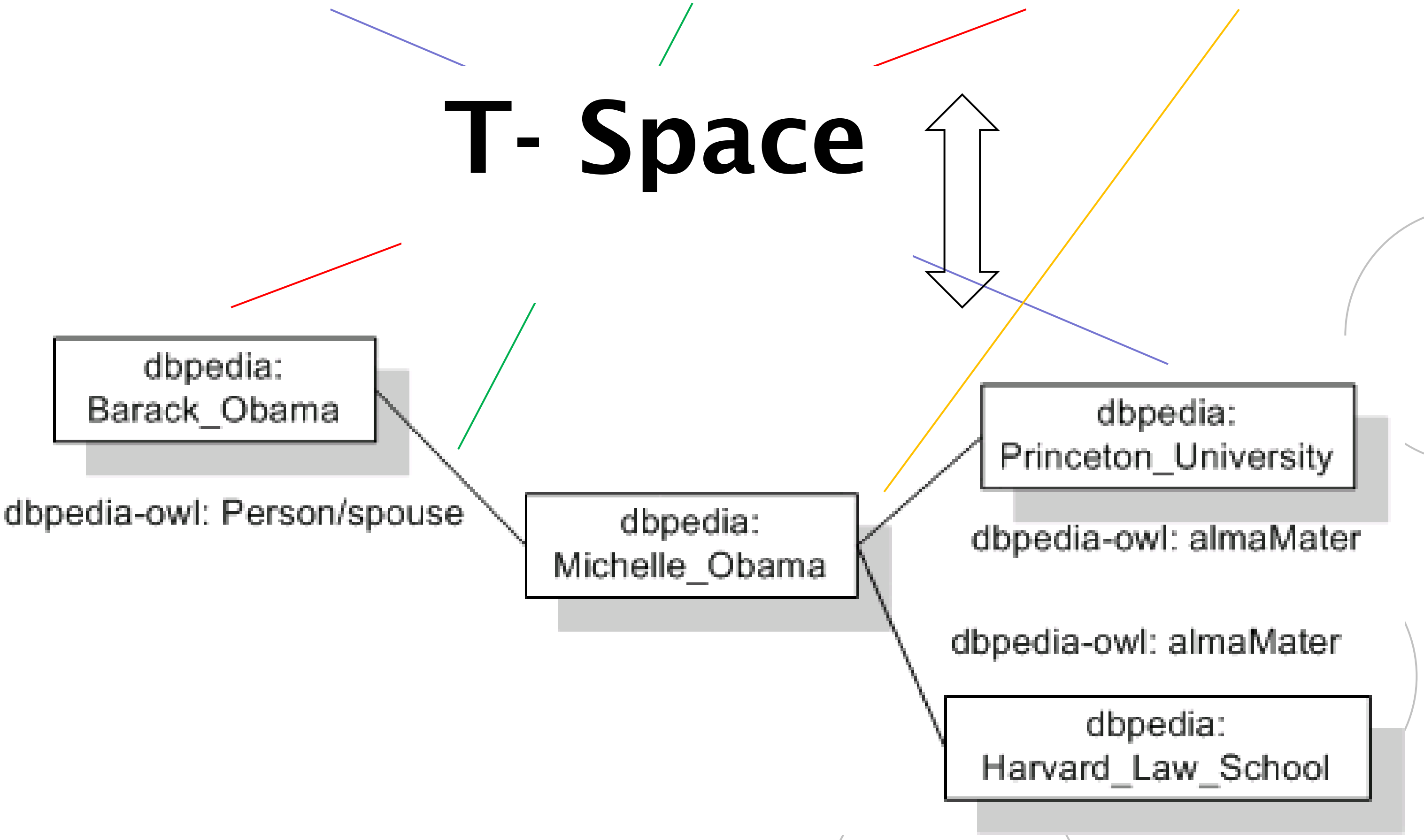
From which university did the wife of Barack Obama graduate?



Structural matching

Semantic Matching Problem

From which university did the wife of Barack Obama graduate?



Strategy

- Best-effort query model (ranked results).
- Use of a distributional semantic model.
- Two phase search process combining *entity search* with *spreading activation search*.

Proposed Approach

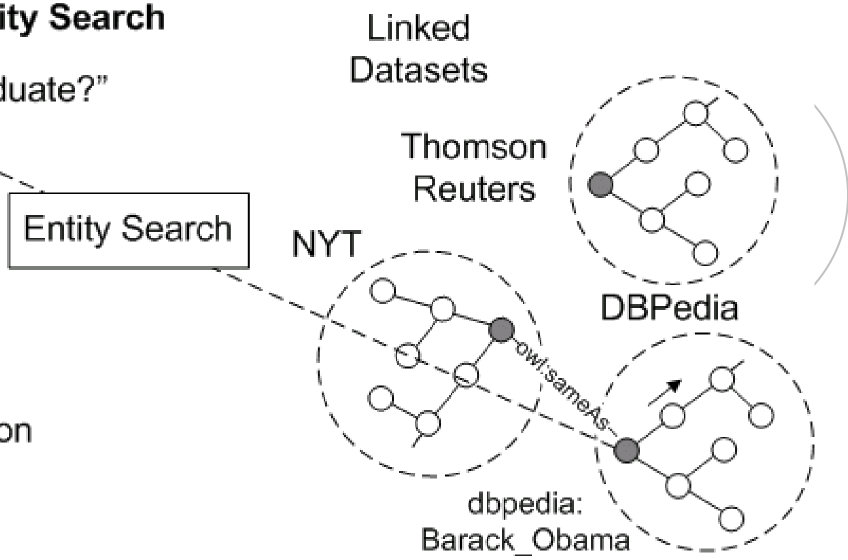
Query Approach Rationale

Query: "From which university did the wife of Barack Obama graduate?"

① **Entity Recognition and Pivot Determination through Entity Search**

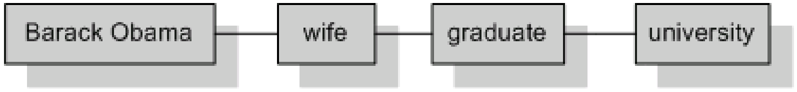
"From which university did the wife of Barack Obama graduate?"

(query pivot entity)



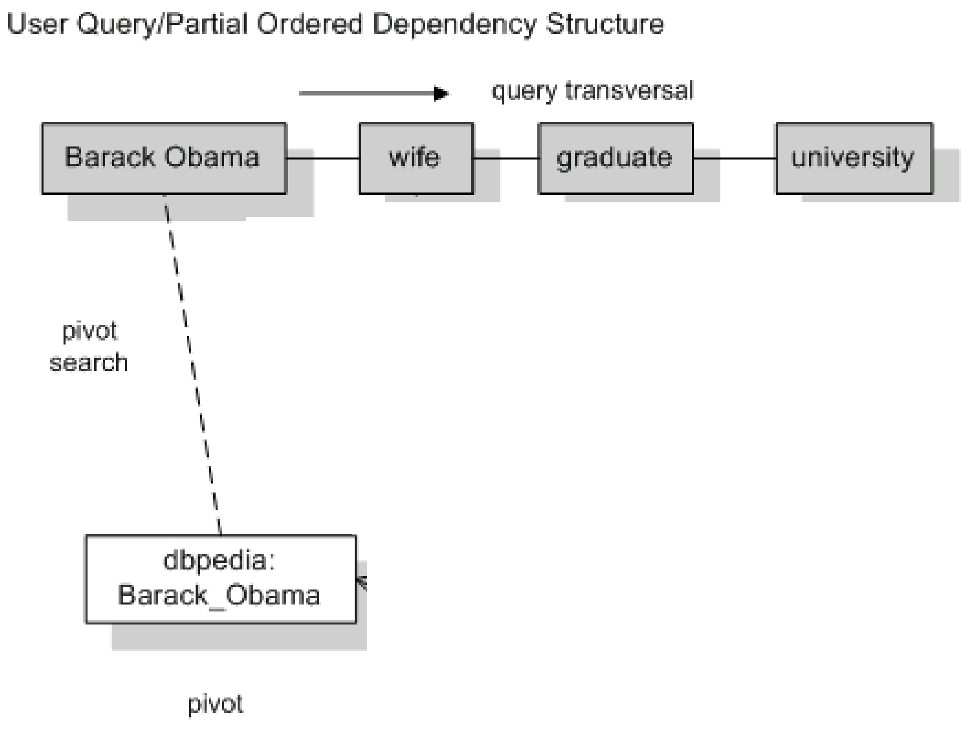
(Pivot entity in the Linked Data Web)

② Query Syntactic Analysis:
Partial Ordered Dependency Structure (PODS) Determination



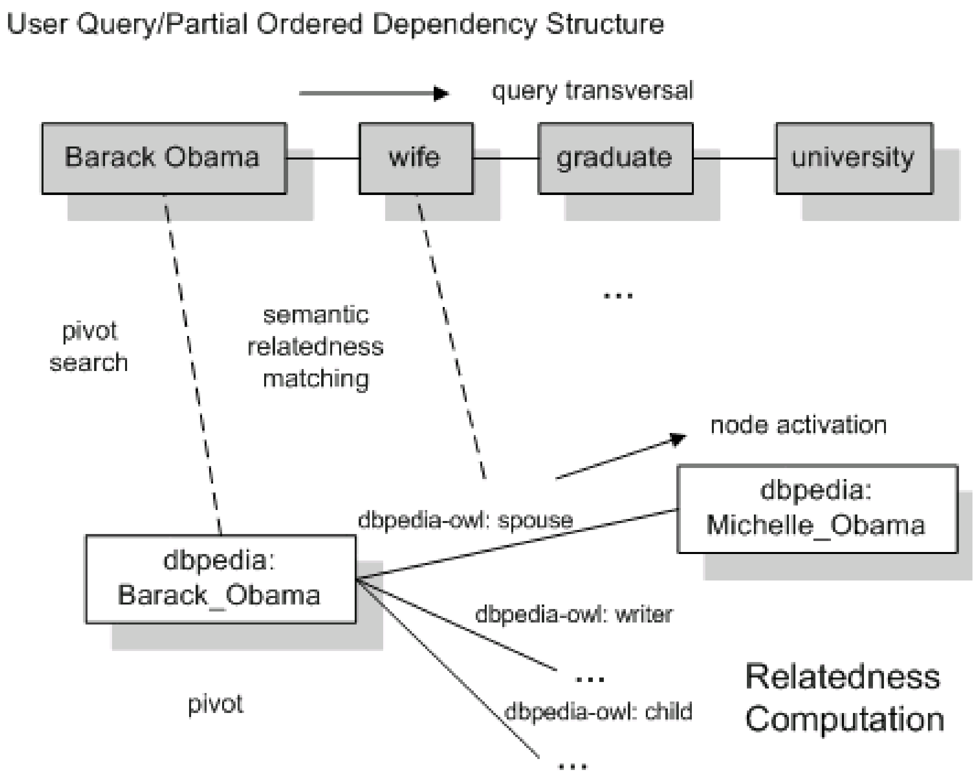
Query Approach Rationale

③ Spreading Activation using Semantic Relatedness



Query Approach Rationale

3 Spreading Activation using Semantic Relatedness



User Query/Partial Ordered Dependency Structure

query transversal

Barack Obama

wife

graduate

university

pivot search

semantic relatedness matching

...

node activation

dbpedia: Michelle_Obama

dbpedia-owl: spouse

dbpedia: Barack_Obama

dbpedia-owl: writer

...

dbpedia-owl: child

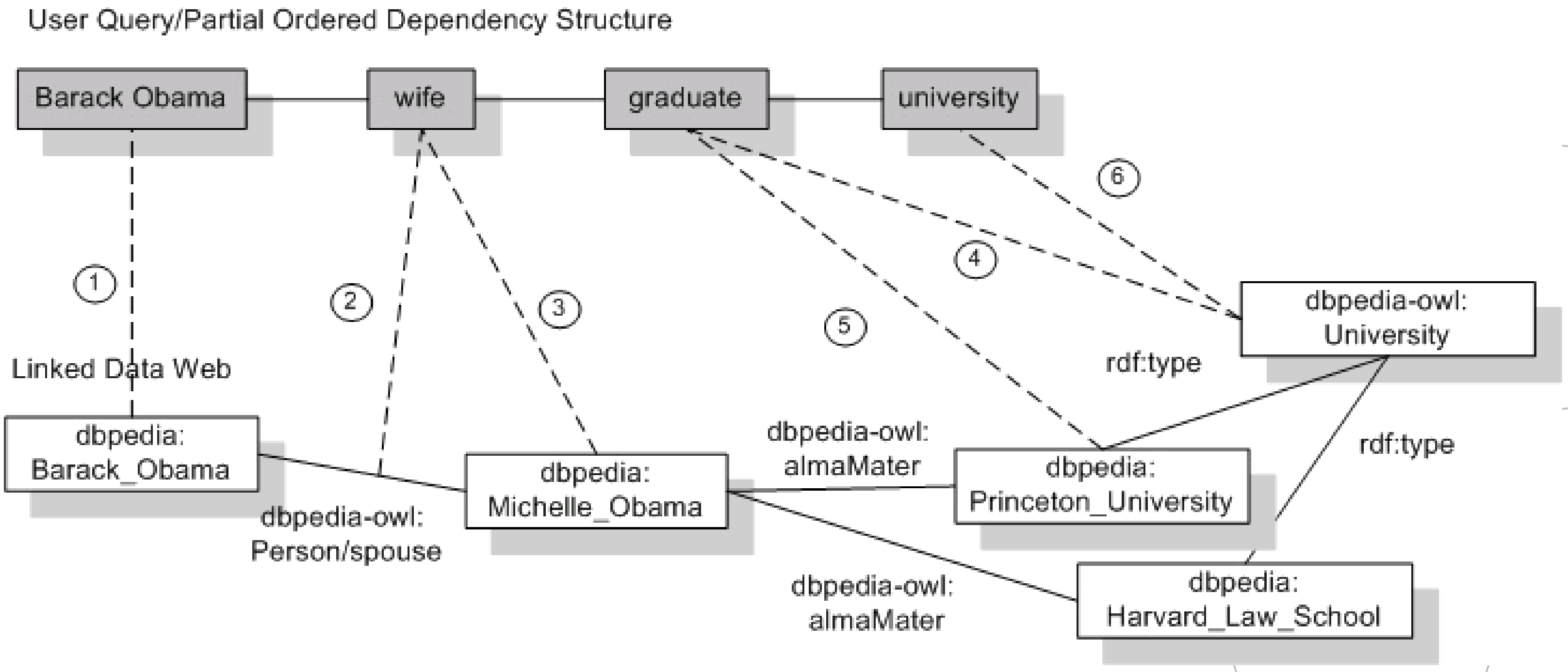
...

pivot

Relatedness Computation

Query Approach Rationale

Final Query- Data Matching:



Semantic Relatedness

- Computation of a measure of “semantic proximity” between two terms.
- Allows a semantic approximate matching between *query terms* and *dataset terms*.
- Most existing approaches use WordNet-based solutions for approximate semantic matching.
- Distributional semantic approaches address these limitations.

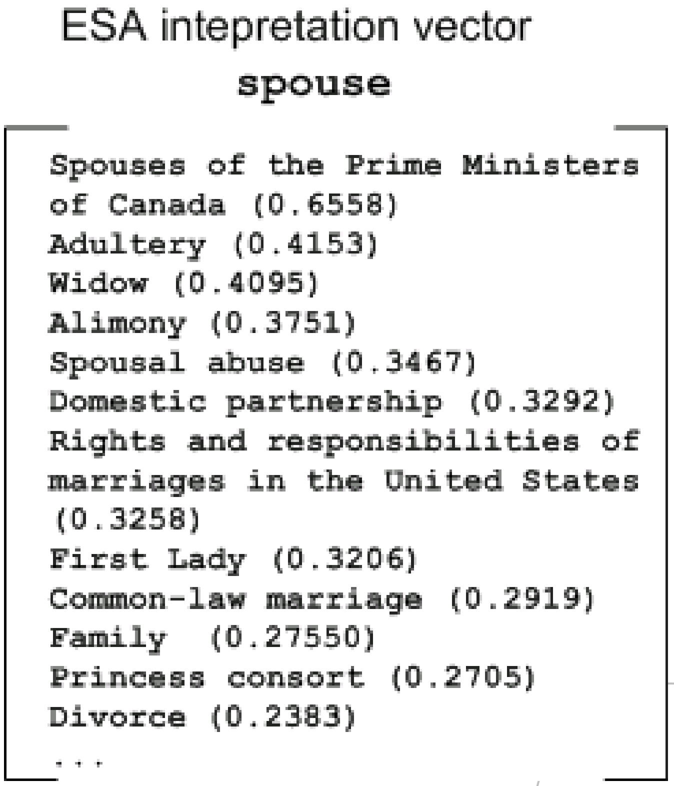
Distributional Semantics

- Assumption: the context surrounding a given word in a text provides important information about its **meaning**.
- Meaning is mediated by word distribution in the corpora.
- Simplified semantic model.

Opera is an art form in which singers and musicians perform a dramatic work combining text (called a libretto) and musical score. **Opera** is part of the Western classical music tradition. **Opera** incorporates many of the elements of spoken theatre, such as acting, scenery, and costumes and sometimes includes dance. The performance is typically given in an opera house, accompanied by an orchestra or smaller musical ensemble.

Explicit Semantic Analysis (ESA)

- Based on Wikipedia.
- Interpretation vector using Wikipedia articles titles.



ESA intepretation vector

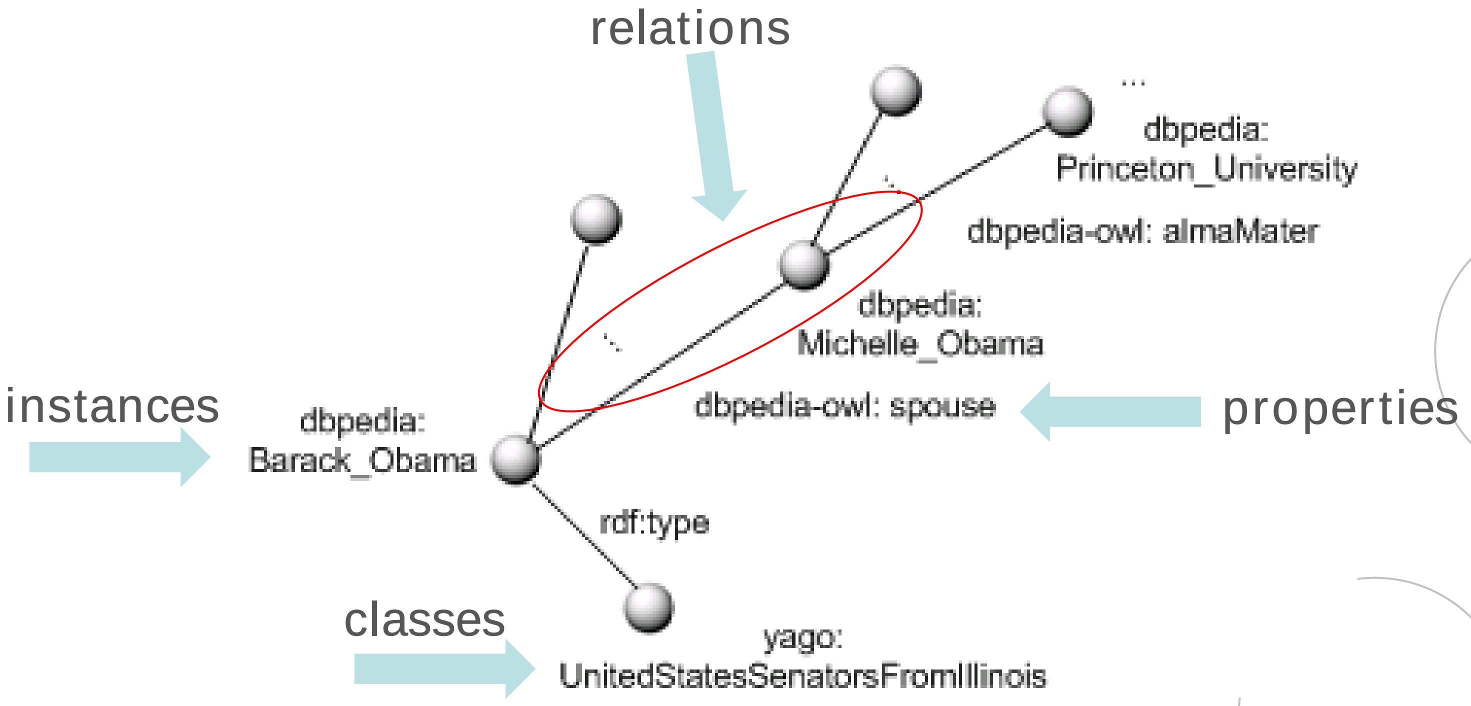
spouse

Spouses of the Prime Ministers of Canada (0.6558)
Adultery (0.4153)
Widow (0.4095)
Alimony (0.3751)
Spousal abuse (0.3467)
Domestic partnership (0.3292)
Rights and responsibilities of marriages in the United States (0.3258)
First Lady (0.3206)
Common-law marriage (0.2919)
Family (0.27550)
Princess consort (0.2705)
Divorce (0.2383)
...

Building the T- Space (Steps)

- Building the distributional semantic model using ESA.
- Construction of instances spaces (TF/IDF).
- Construction of classes spaces (ESA).
- Construction of relation spaces (ESA).

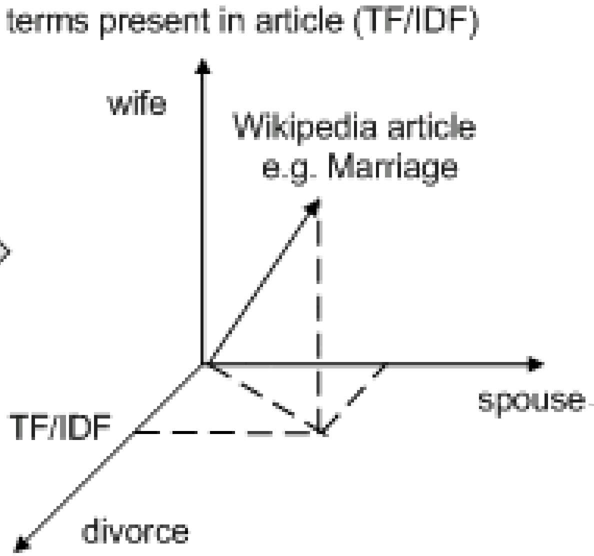
Building the T- Space



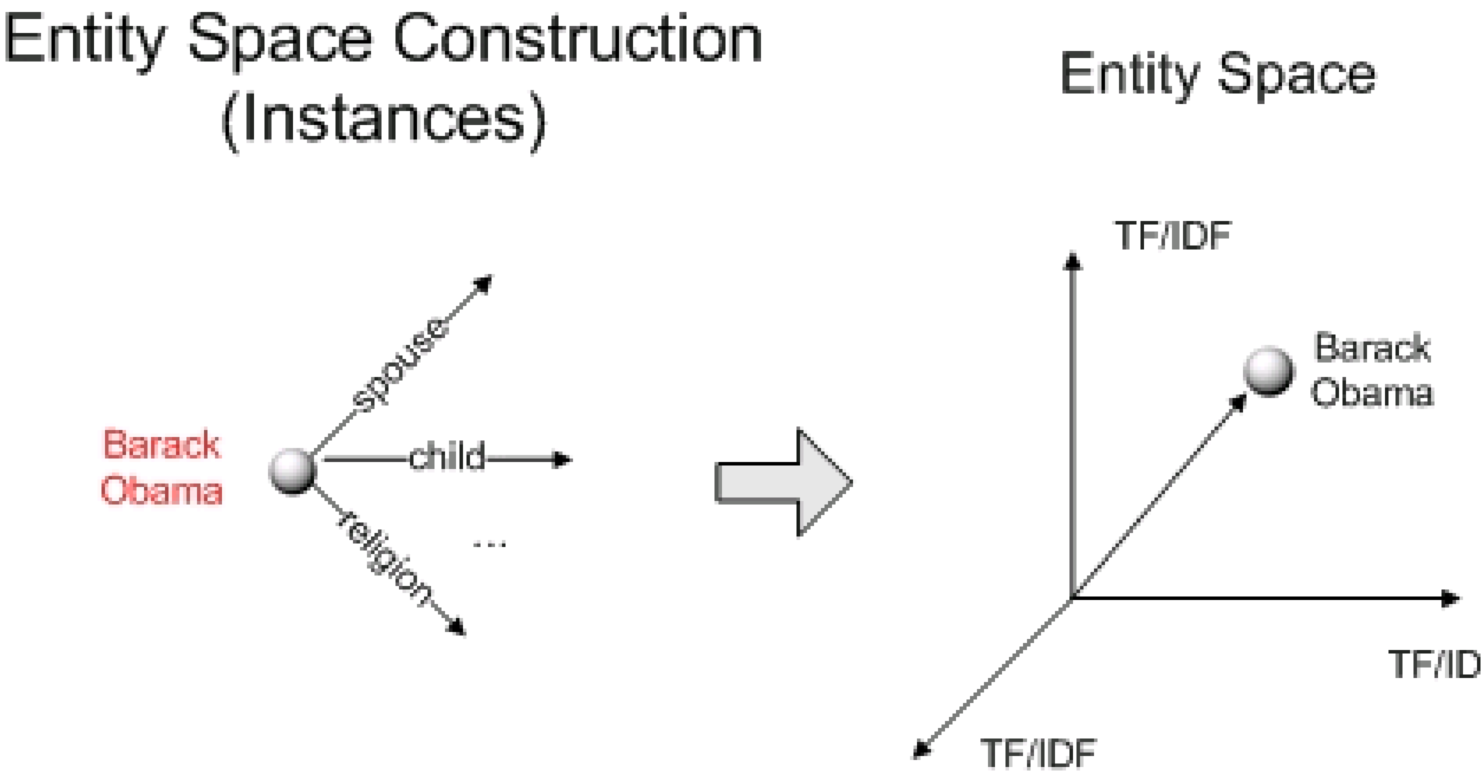
Building the T- Space

Universal ESA Space Construction

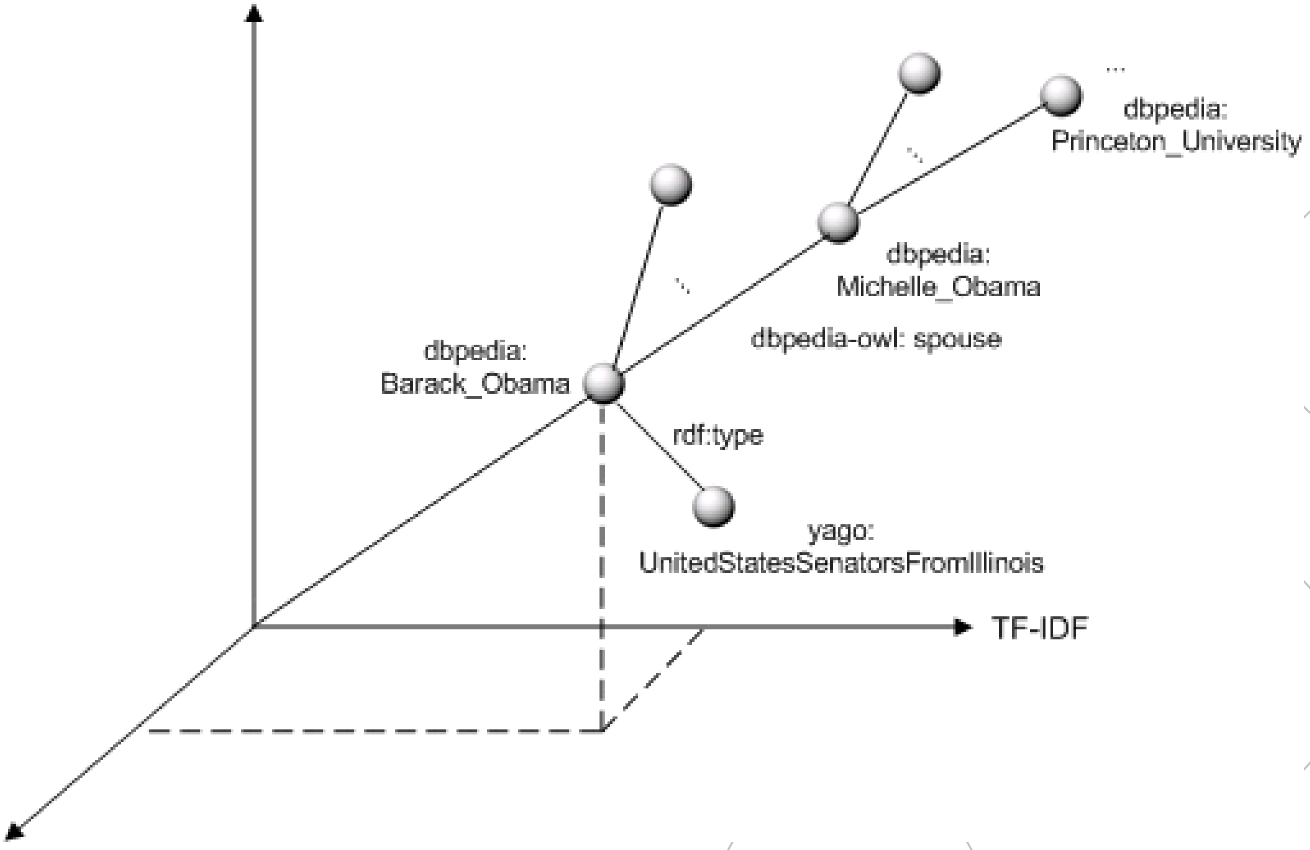
Universal ESA Concept Space


terms present in article (TF/IDF)
wife
Wikipedia article e.g. Marriage
spouse
TF/IDF
divorce

Instances


Entity Space Construction
(Instances)
Barack
Obama
spouse
child
religion
...
Entity Space
TF/IDF
Barack
Obama
TF/IDF
TF/IDF

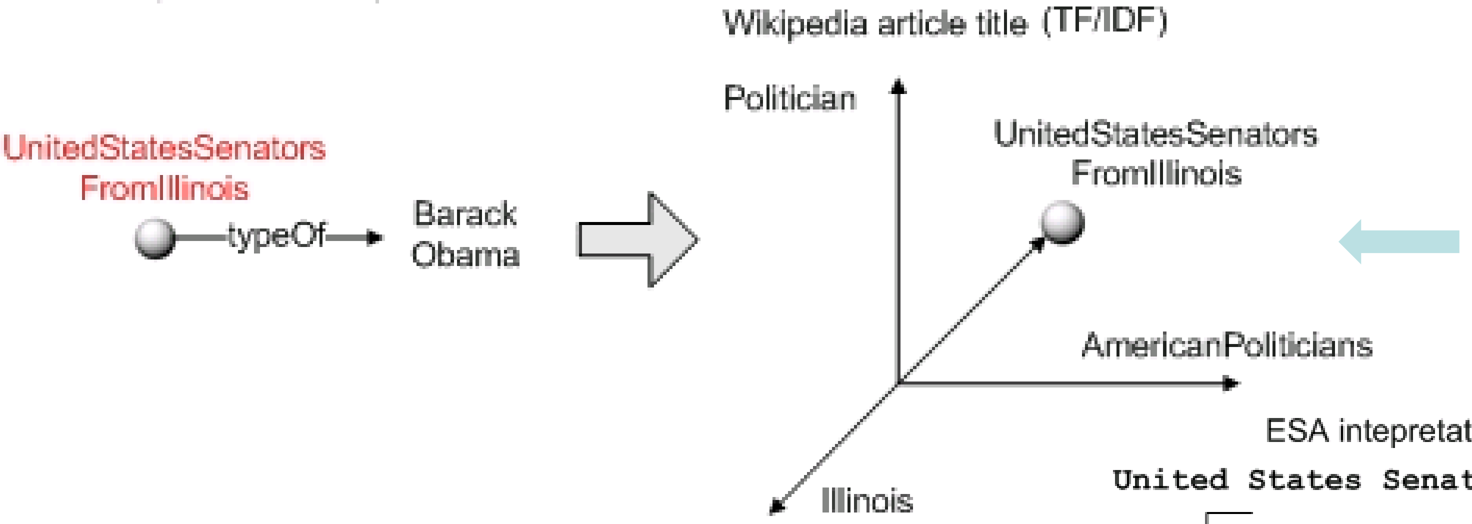
Building the T- Space



Classes

Entity Space Construction (Classes)

Entity Space


Wikipedia article title (TF/IDF)
UnitedStatesSenators FromIllinois
typeOf
Barack Obama
Politician
UnitedStatesSenators FromIllinois
AmericanPoliticians
Illinois

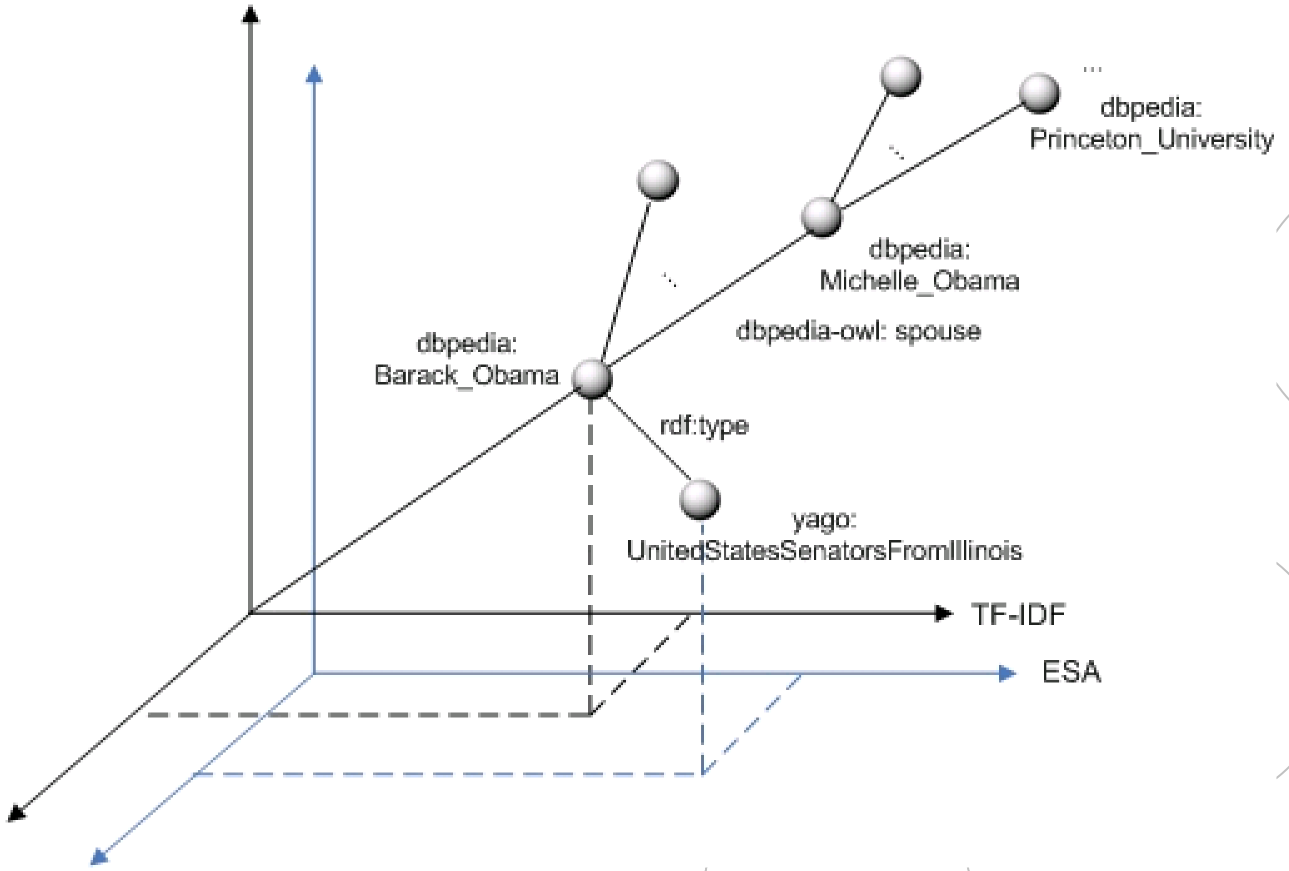
Universal ESA Space

ESA intepretation vector

United States Senators From Illinois

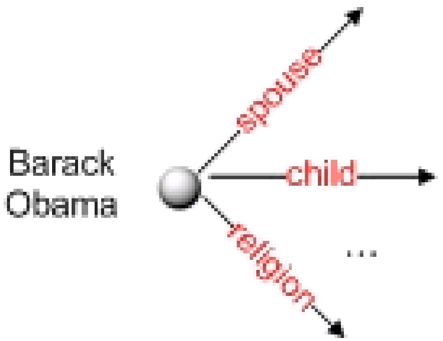
```
Governor of Illinois (0.1219)
Senate  (0.10695)
History of Illinois (0.09895)
Western Illinois University
(0.09180)
Springfield, Illinois (0.09122)
University of Illinois  (0.09106)
Normal, Illinois (0.09028)
Illinois Central Railroad
(0.08792)
Illinois (0.08704)
Peter Fitzgerald (0.08685)
Cairo, Illinois (0.08670)
David Davis (senator) (0.08596)
Carbondale, Illinois (0.08548)
Illinois State University
(0.084508)
...
```

Building the T- Space

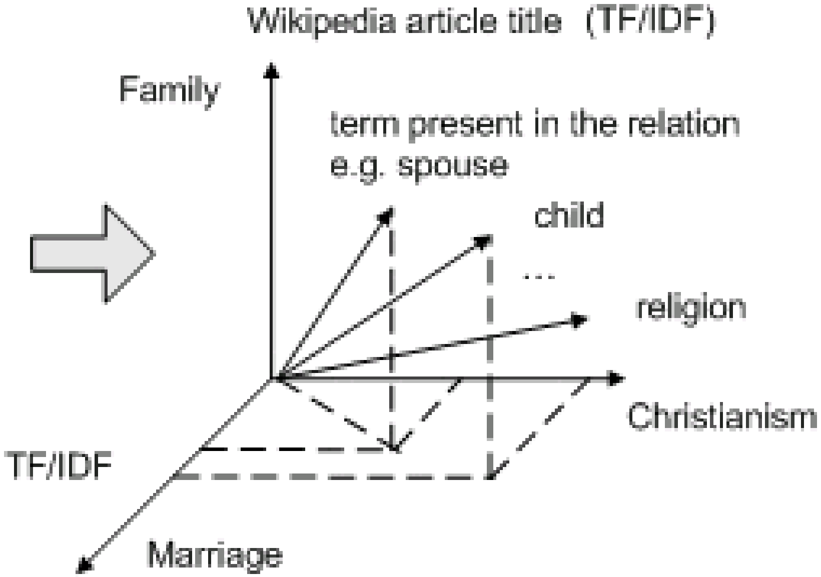


Relations

Relation Subspaces Construction



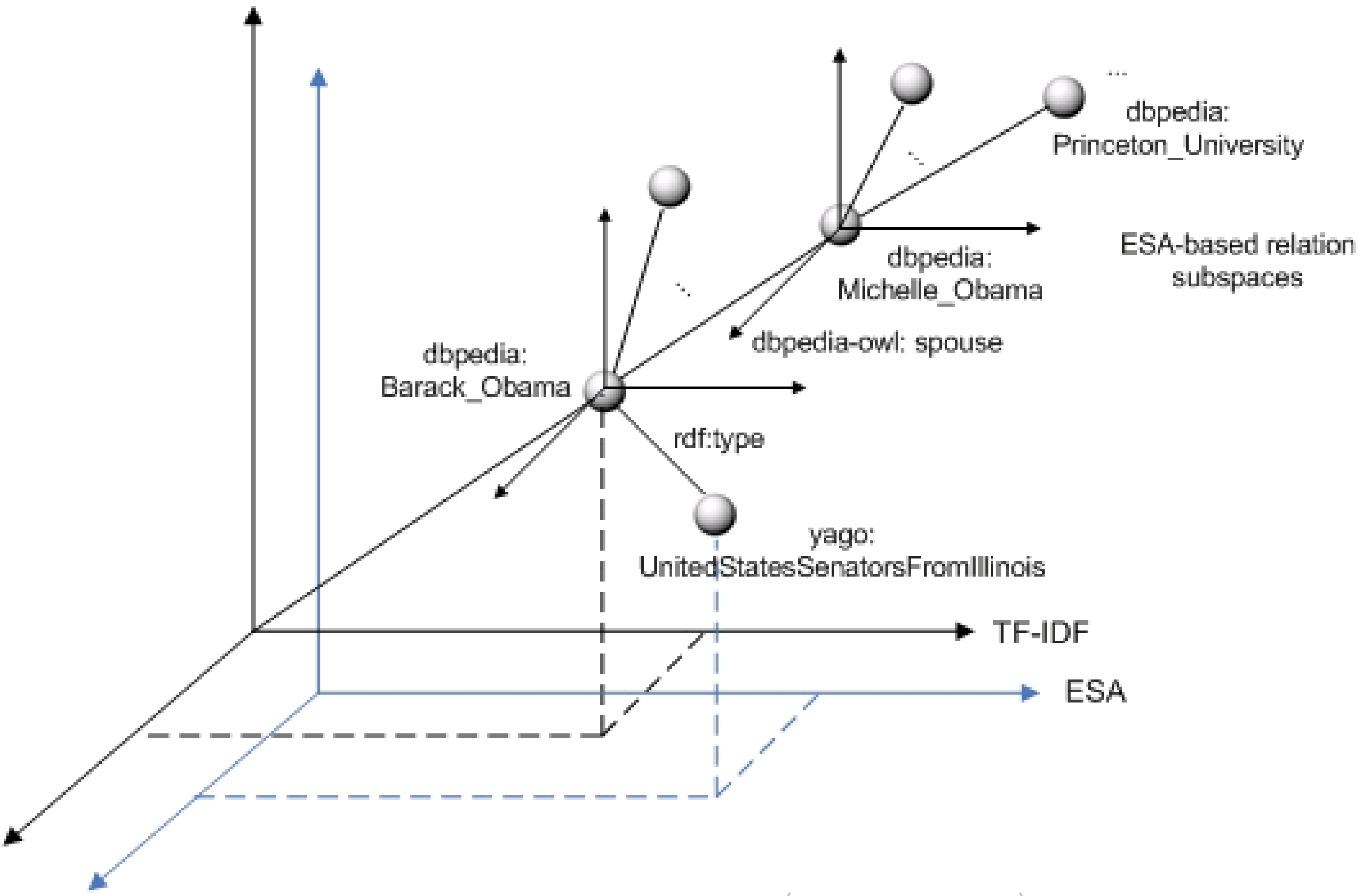
Relation subspace ESA-based index


Wikipedia article title (TF/IDF)
Family
term present in the relation e.g. spouse
child
...
religion
Christianism
TF/IDF
Marriage

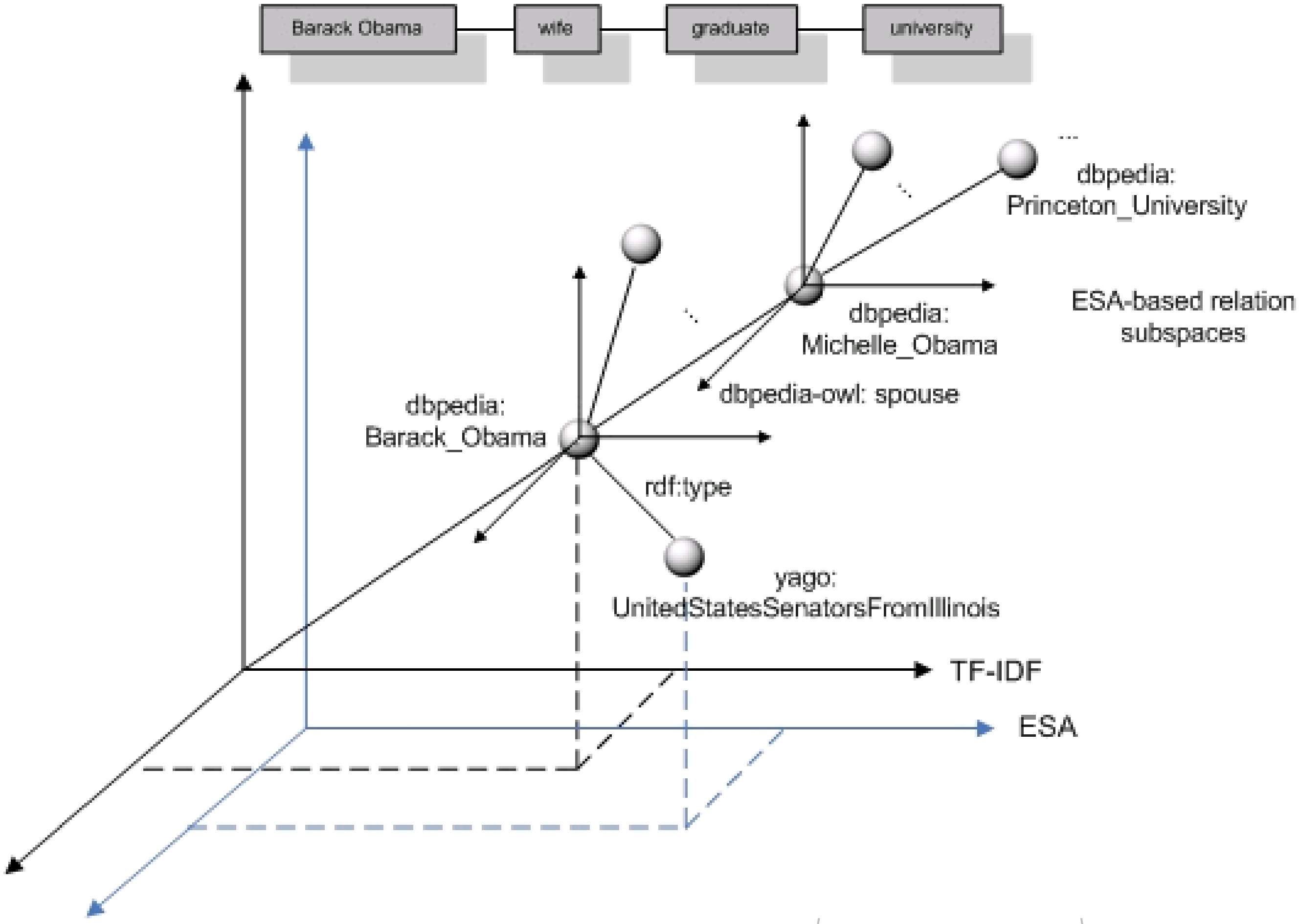
ESA intepretation vector spouse

Spouses of the Prime Ministers of Canada (0.6558)
Adultery (0.4153)
Widow (0.4095)
Alimony (0.3751)
Spousal abuse (0.3467)
Domestic partnership (0.3292)
Rights and responsibilities of marriages in the United States (0.3258)
First Lady (0.3206)
Common-law marriage (0.2919)
Family (0.27550)
Princess consort (0.2705)
Divorce (0.2383)
...

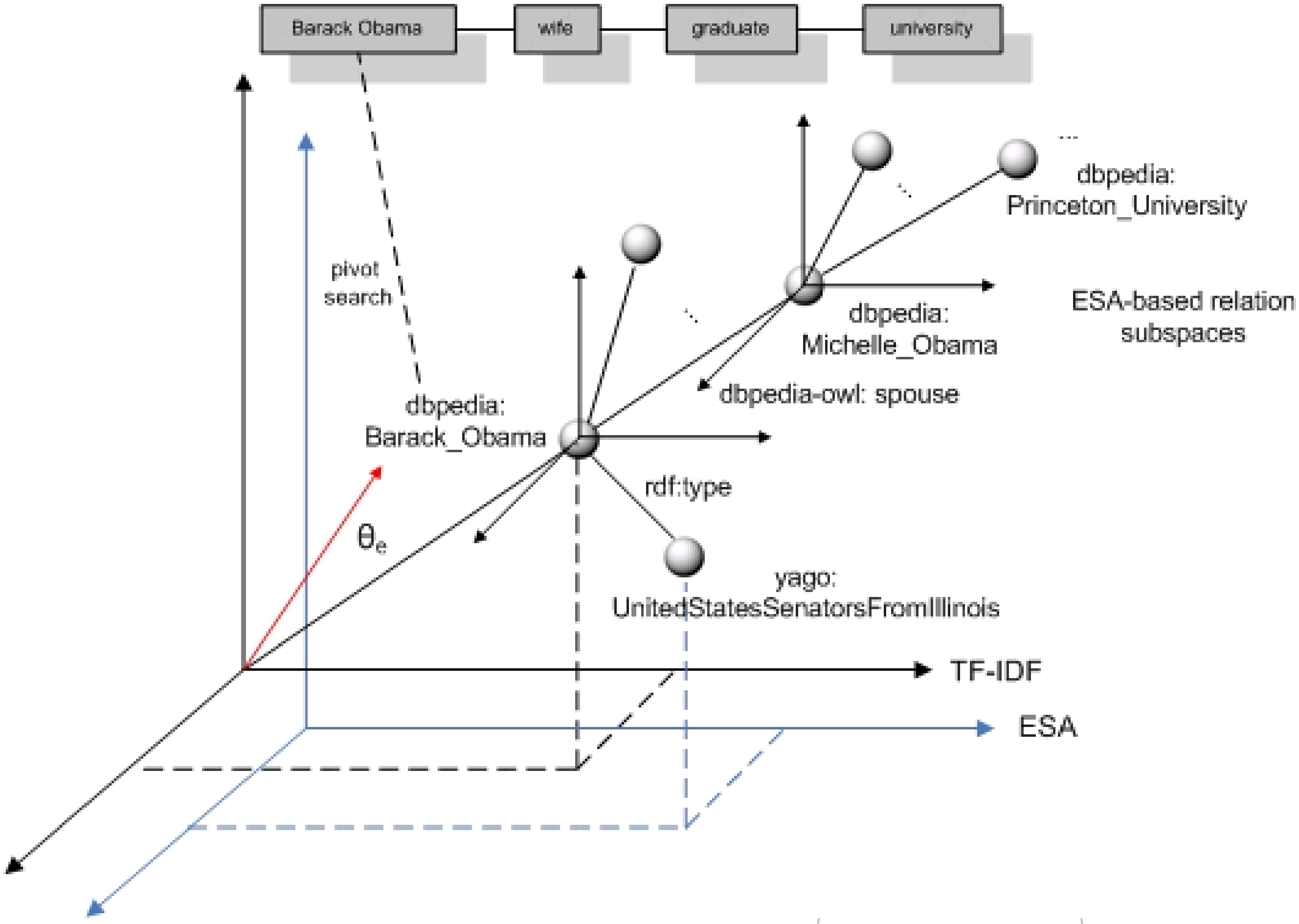
Building the T- Space



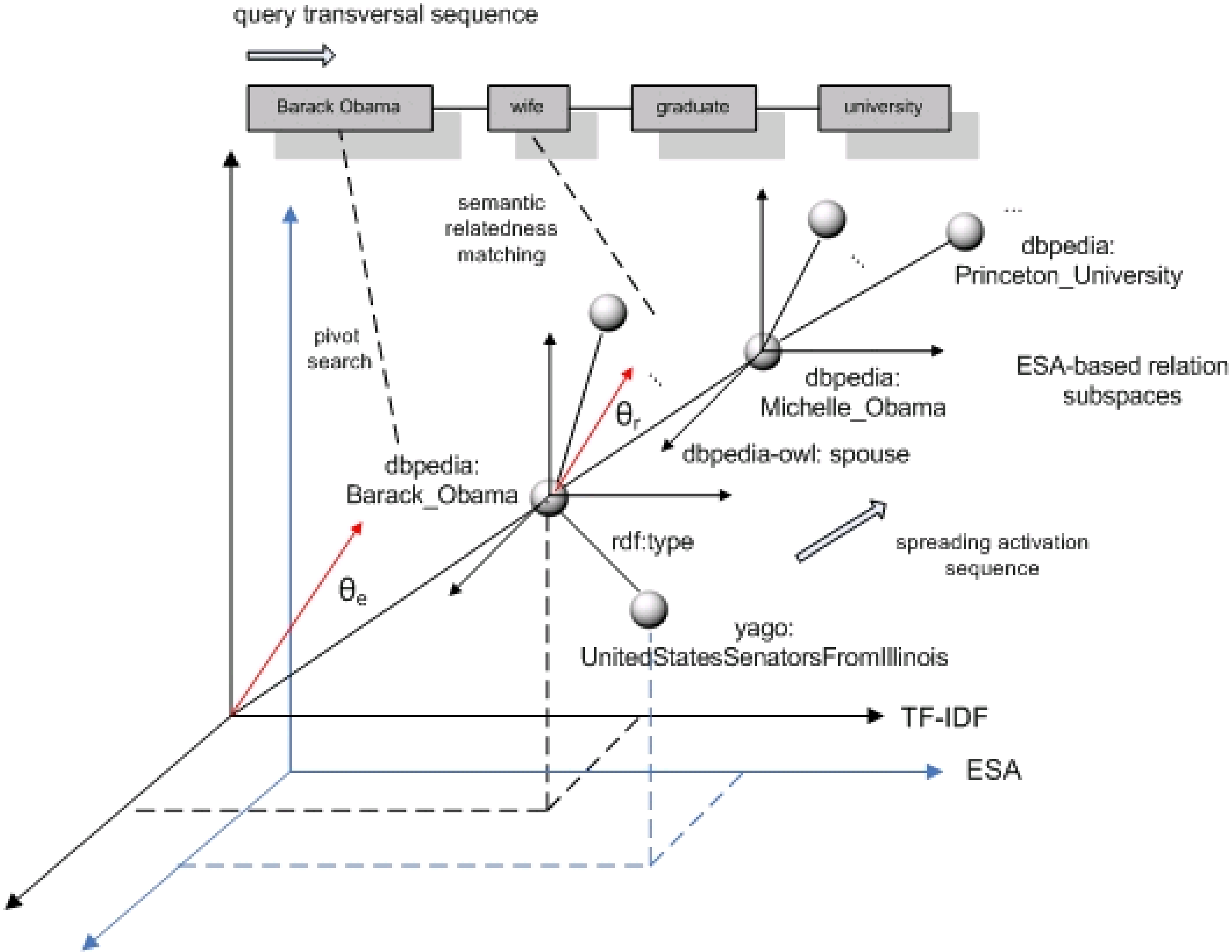
Querying the T- Space



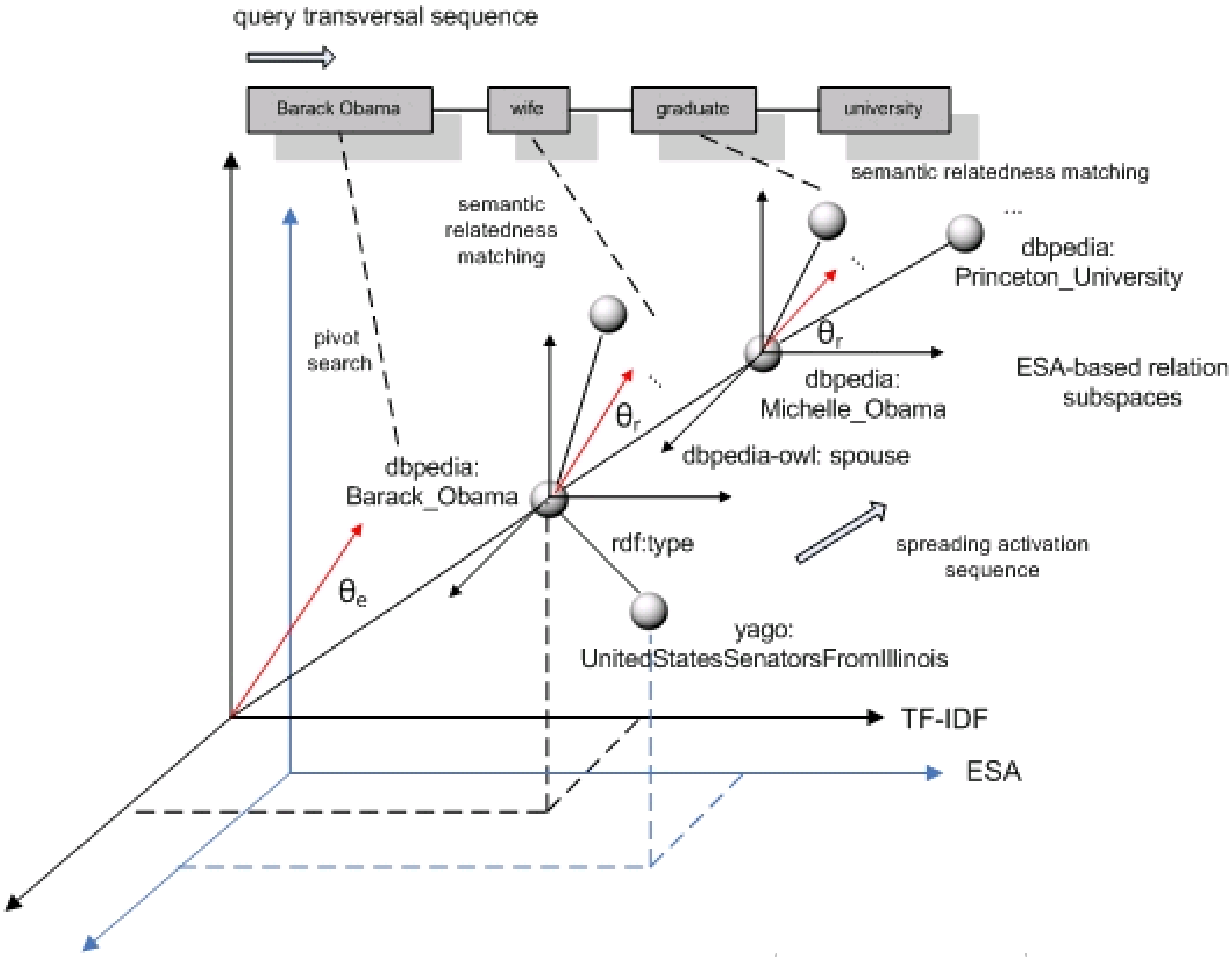
Querying the T- Space



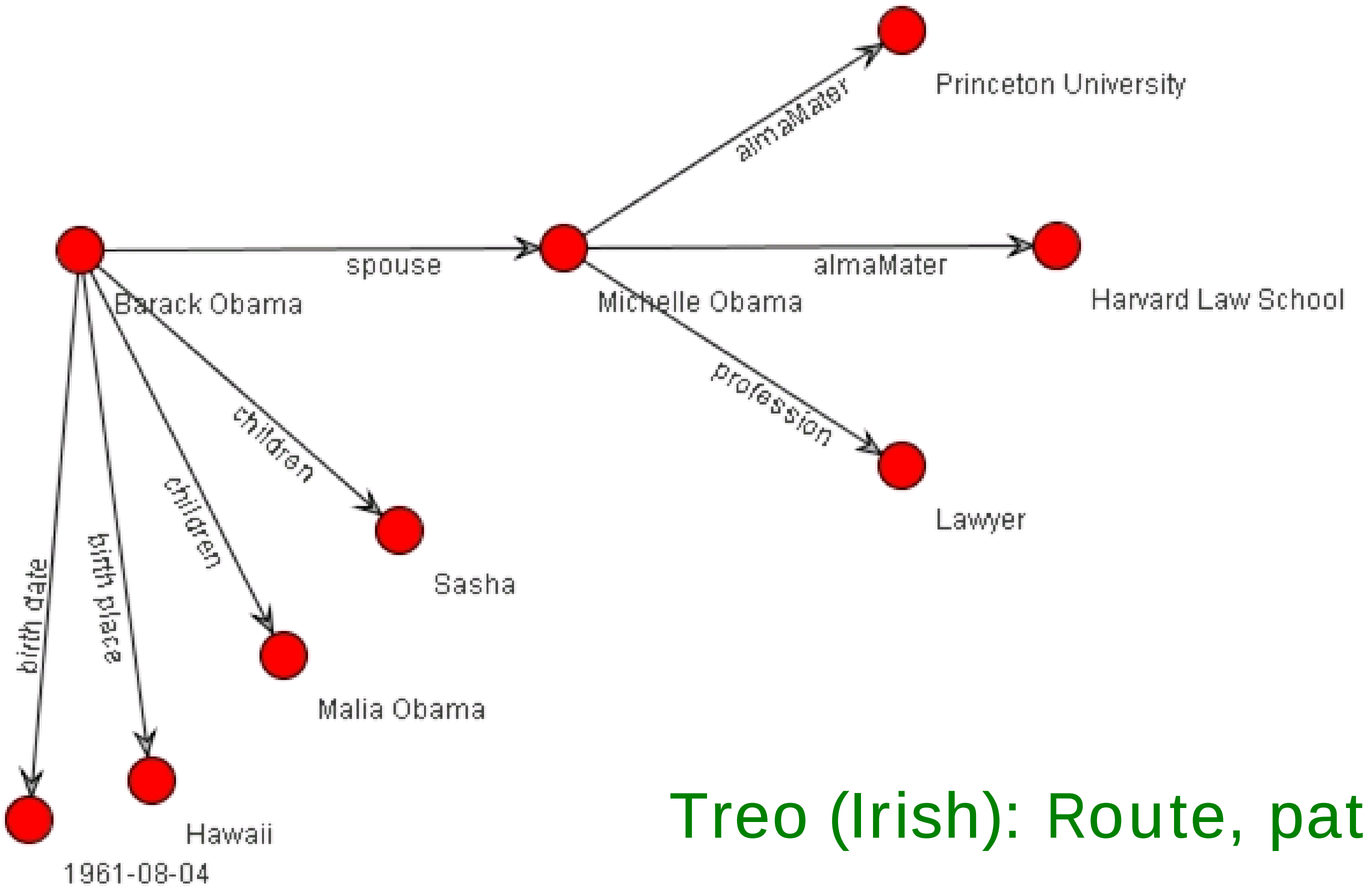
Querying the T- Space


query transversal sequence
Barack Obama
wife
graduate
university
semantic relatedness matching
pivot search
dbpedia: Barack_Obama
θ_e
θ_r
dbpedia: Michelle_Obama
dbpedia-owl: spouse
dbpedia: Princeton_University
...
ESA-based relation subspaces
rdf:type
spreading activation sequence
yago: UnitedStatesSenatorsFromIllinois
TF-IDF
ESA

Querying the T- Space


query transversal sequence
Barack Obama
wife
graduate
university
semantic relatedness matching
semantic relatedness matching
...
dbpedia: Princeton_University
pivot search
θr
θr
dbpedia: Michelle_Obama
ESA-based relation subspaces
dbpedia-owl: spouse
dbpedia: Barack_Obama
rdf:type
spreading activation sequence
θe
yago: UnitedStatesSenatorsFromIllinois
TF-IDF
ESA

Treo



Treo (Irish): Route, path

Evaluation

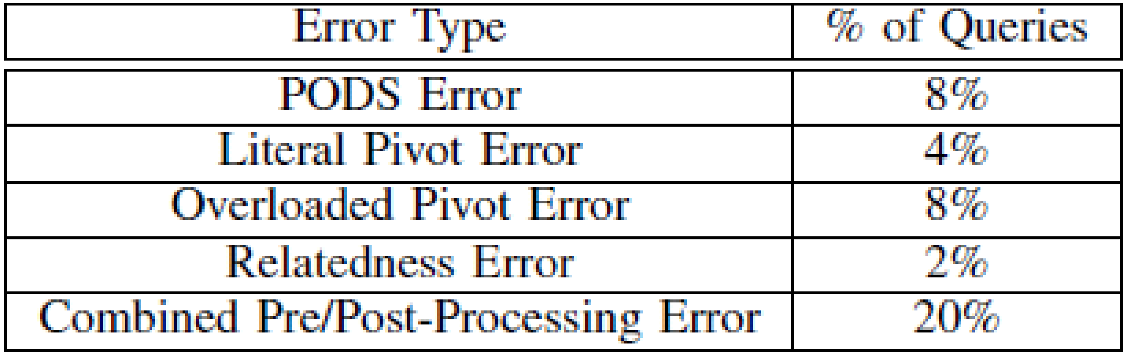
Quality of Results

- QALD DBPedia Training Set.
- 50 natural language queries.
- DBpedia 3.6.

Full DBPedia QuerySet (50 queries)			
Avg. Precision	Avg. Recall	MRR	% of queries answered
0.482	0.491	0.516	58%

Partial DBPedia QuerySet (38 queries)			
Avg. Precision	Avg. Recall	MRR	% of queries answered
0.634	0. 645	0.679	76%

Error Distribution



Error Type	% of Queries
PODS Error	8%
Literal Pivot Error	4%
Overloaded Pivot Error	8%
Relatedness Error	2%
Combined Pre/Post-Processing Error	20%

Conclusion & Future Work

- The T-Space semantic model shows a promising direction for providing data model independent queries over RDF data.
- Improvement of *semantic tractability*.
- The distributional semantic model supports a flexible matching between query terms and dataset terms in a best-effort scenario.
- Further improvements are needed:
  - □ QA features (e.g. answer type detection, operators).
  - □ User feedback mechanisms (disambiguation).
  - □ Entity recognition for complex classes.