

Creating and Exploiting a Web of Semantic Data

Tim Finin

University of Maryland, Baltimore County

joint work with Zareen Syed (UMBC) and
colleagues at the Johns Hopkins University Human
Language Technology Center of Excellence

ICAART 2010, 24 January 2010

Overview

- Conclusion
- Introduction
- A Web of linked data
- Wikitology
- Applications
- Conclusion

Conclusion

- The Web has made people *smarter* and more *capable*, providing easy access to the world's knowledge and services
- Software agents need better access to a Web of data and knowledge to enhance their intelligence
- Some key technologies are ready to exploit: Semantic Web, linked data, RDF search engines, DBpedia, Wikitology, information extraction, etc.

The Age of Big Data

- Massive amounts of data is available today on the Web, both for people and agents
- This is what's driving Google, Bing, Yahoo
- Human language advances also driven by availability of unstructured data, text & speech
- Large amounts of structured & semi-structured data is also coming online, including RDF
- We can exploit this data to enhance our intelligent agents and services

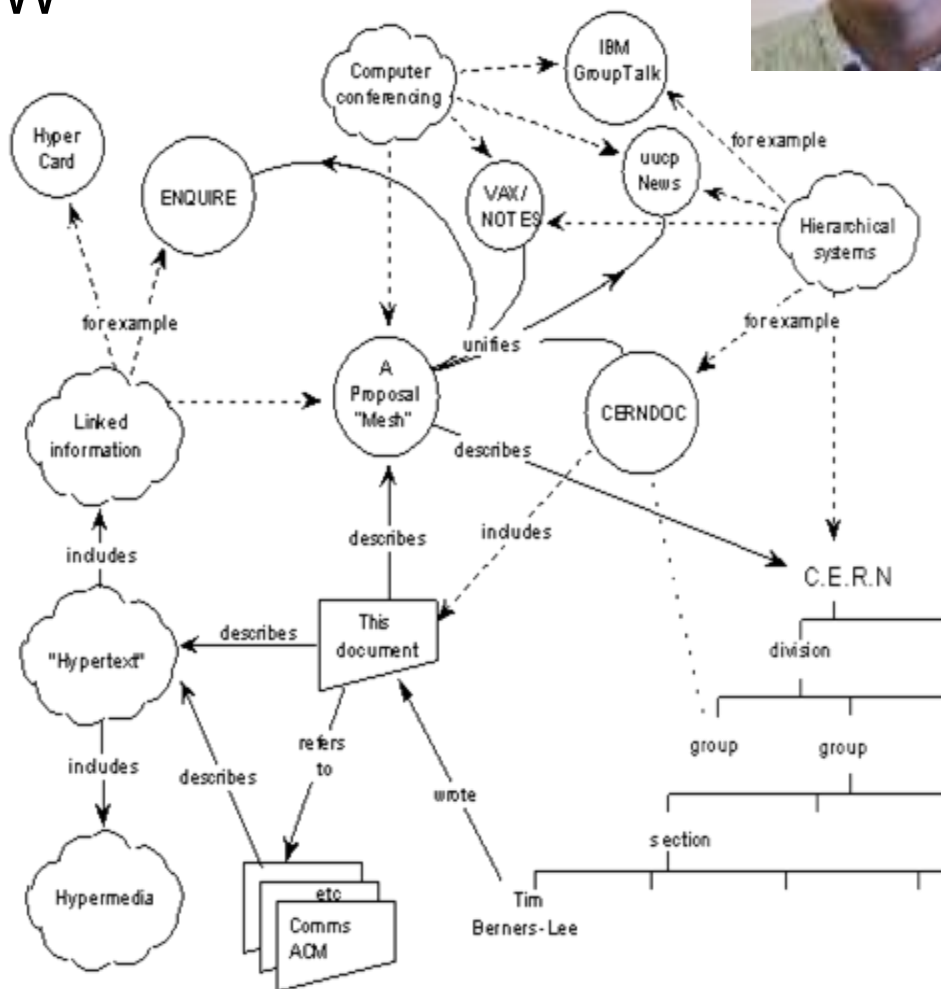
Twenty years ago...



Tim Berners-Lee's 1989 WWW proposal described a web of relationships among named objects unifying many info. management tasks.

Capsule history

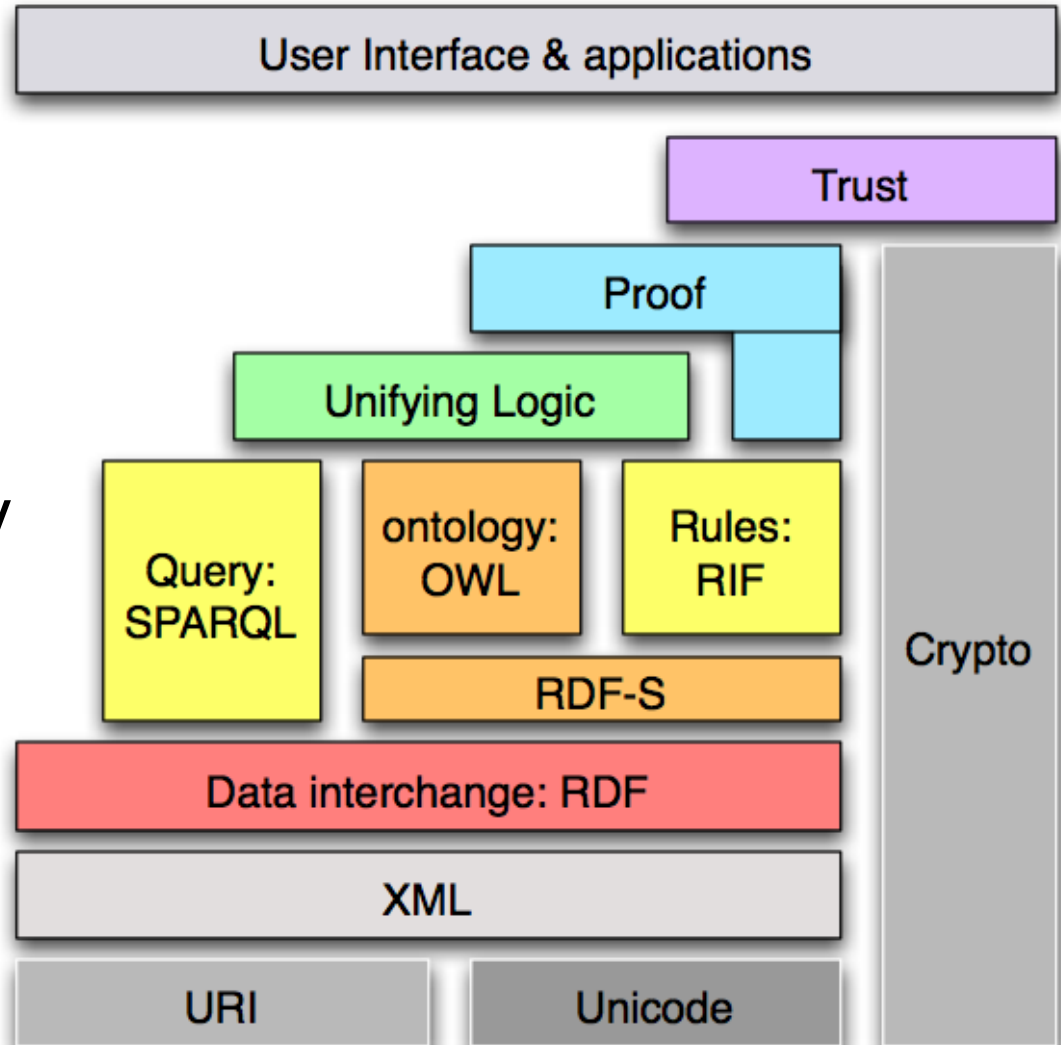
- Guha's MCF (~94)
- XML+MCF=>RDF (~96)
- RDF+OO=>RDFS (~99)
- RDFS+KR=>DAML+OIL (00)
- W3C's SW activity (01)
- W3C's OWL (03)
- SPARQL, RDFa (08)



<http://www.w3.org/History/1989/proposal.html>

Ten yeas ago...

- The W3C began developing standards to support the Semantic Web
- The vision, technology and use cases are still evolving
- Moving from a Web of documents to a Web of data



Today's LOD Cloud

The diagram illustrates a dense network of Linked Open Data (LOD) endpoints, categorized by color and interconnected by numerous relationships. The central hub is **DBpedia** (cyan), which connects to a wide range of datasets. Other prominent hubs include **GeoNames** (yellow), **DBLP RKB Explorer** (green), and **UniProt** (pink). The network is organized into several clusters:

- Music and Media (Blue):** Includes BBC Music, MySpace Wrapper, Music-brainz, Audio-Scrobbler, Jamendo, BBC Later + TOTP, BBC John Peel, BBC Playcount Data, and BBC Programmes.
- General Web and Geospatial (Yellow):** Includes Geo-names, Pub Guide, riese, US Census Data, Linked GeoData, Gov-Track, and World Fact-book.
- Research and Reference (Cyan):** Includes W3C WordNet, UMBEL, Yago, Open Cyc, and linkedMDB.
- Academic and Scientific (Green):** Includes ACM, DBLP RKB Explorer, eprints, IEEE, CiteSeer, DBLP Hannover, DBLP Berlin, Freebase, Open Calais, flickr wrapper, Project Gutenberg, Euro-stat, Crunch Base, FOAF profiles, SIOC Sites, Revyu, Virtuoso Sponger, RKB ECS South-ampton, RAE 2001, National Science Foundation, CORDIS, and Newcastle.
- Biological and Medical (Pink):** Includes UniRef, UniParc, PROSITE, Taxonomy, UniProt, Pfam, ProDom, PDB, Inter Pro, Gene Ontology, ChEBI, OMIM, HGNC, MGI, PubMed, GeneID, Reactome, UniParc, UniRef, and many others.

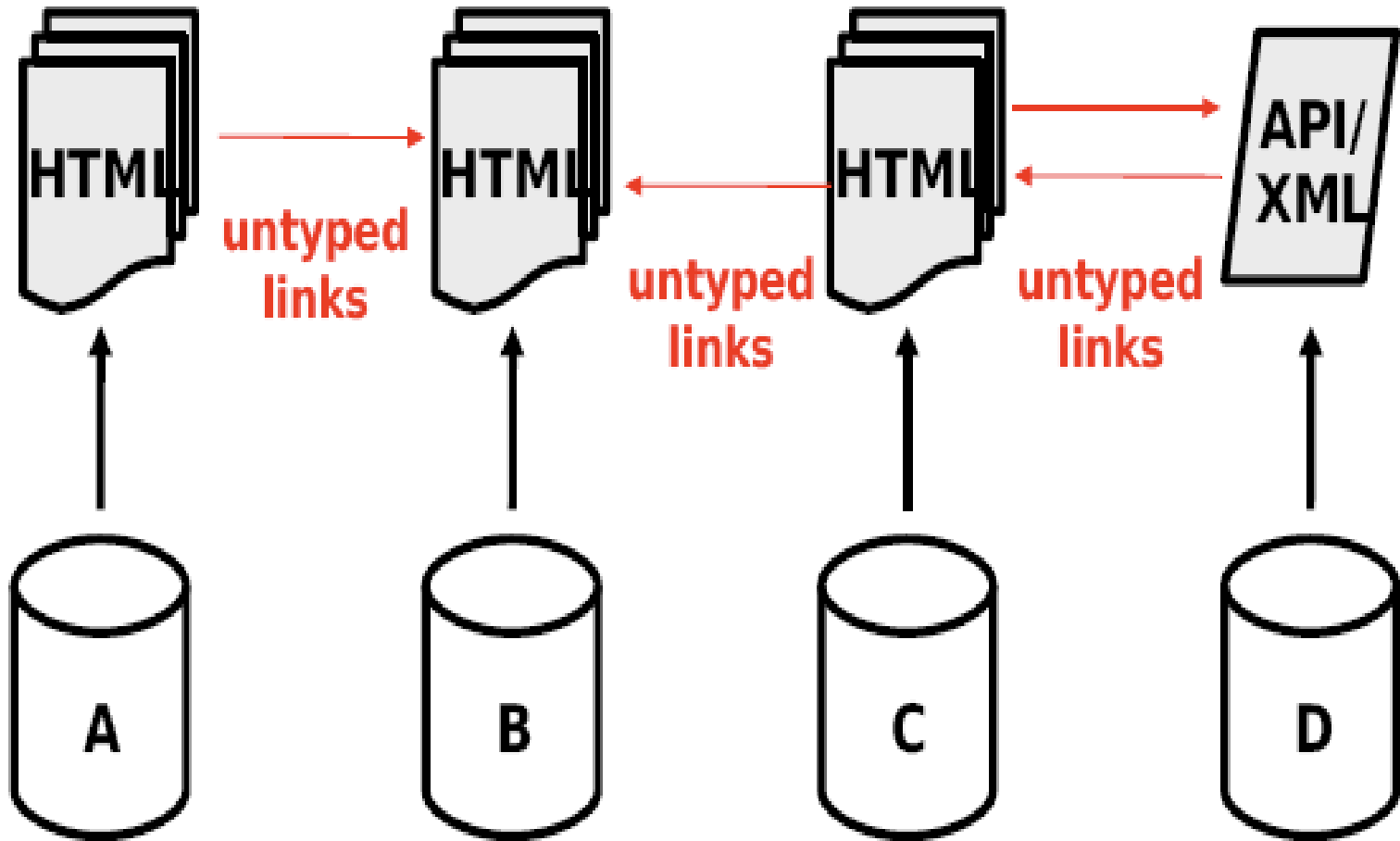
The diagram demonstrates the vast and interconnected nature of the LOD ecosystem, showing how data from various domains is linked together, enabling a rich and diverse web of information.

introduction • linked data • wikipitology • applications • conclusion

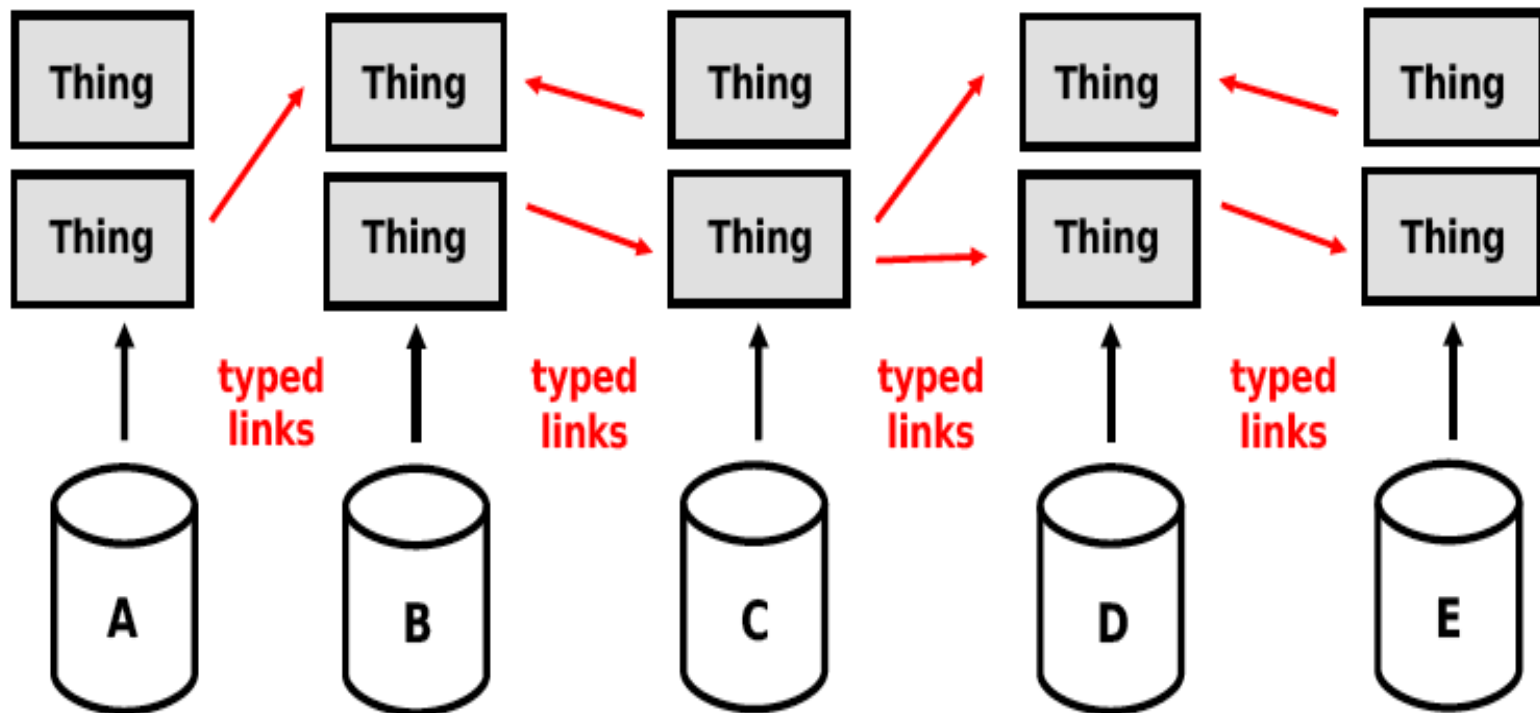
Today's LOD Cloud

- ~5B integrated facts published on Web as RDF
- ~100 datasets
- Arcs represent “joins” across datasets
- Available to download or query via public SPARQL servers
- Updated and improved periodically

From a Web of documents



To a Web of (Linked) Data

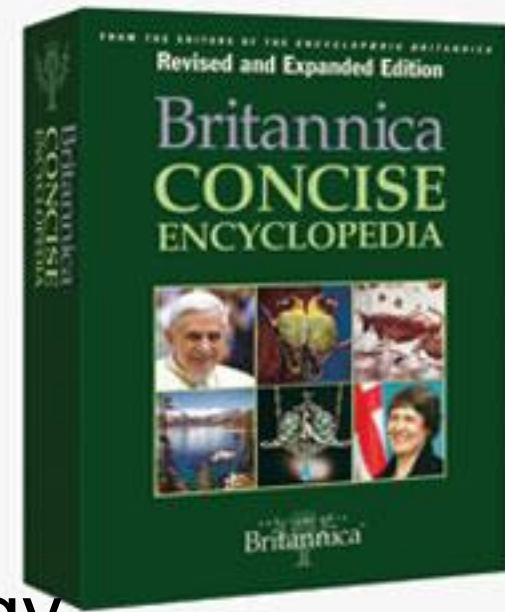


Wikipedia, DBpedia and inked data

- **Wikipedia** as a source of knowledge
 - Wikis have turned out to be great ways to collaborate on building up knowledge resources
- Wikipedia as an **ontology**
 - Every Wikipedia page is a concept or object
- Wikipedia as **RDF data**
 - Map this ontology into RDF
- DBpedia as the lynchpin for **Linked Data**
 - Exploit its breadth of coverage to integrate things

Wikipedia is the new Cyc

- There's a history of using encyclopedias to develop KBs
- Cyc's original goal (c. 1984) was to encode the knowledge in a desktop encyclopedia
- And use it as an integrating ontology
- Wikipedia is comparable to Cyc's original desktop encyclopedia
- But it's machine accessible and malleable
- And available (mostly) in RDF!



Dbpedia: Wikipedia in RDF



- A community effort to extract **structured information** from Wikipedia and publish as RDF on the Web
- Effort started in 2006 with EU funding
- Data and software open sourced
- DBpedia doesn't extract information from Wikipedia's text (yet), but from its structured information, e.g., infoboxes, links, categories, redirects, etc.

DBpedia's ontologies

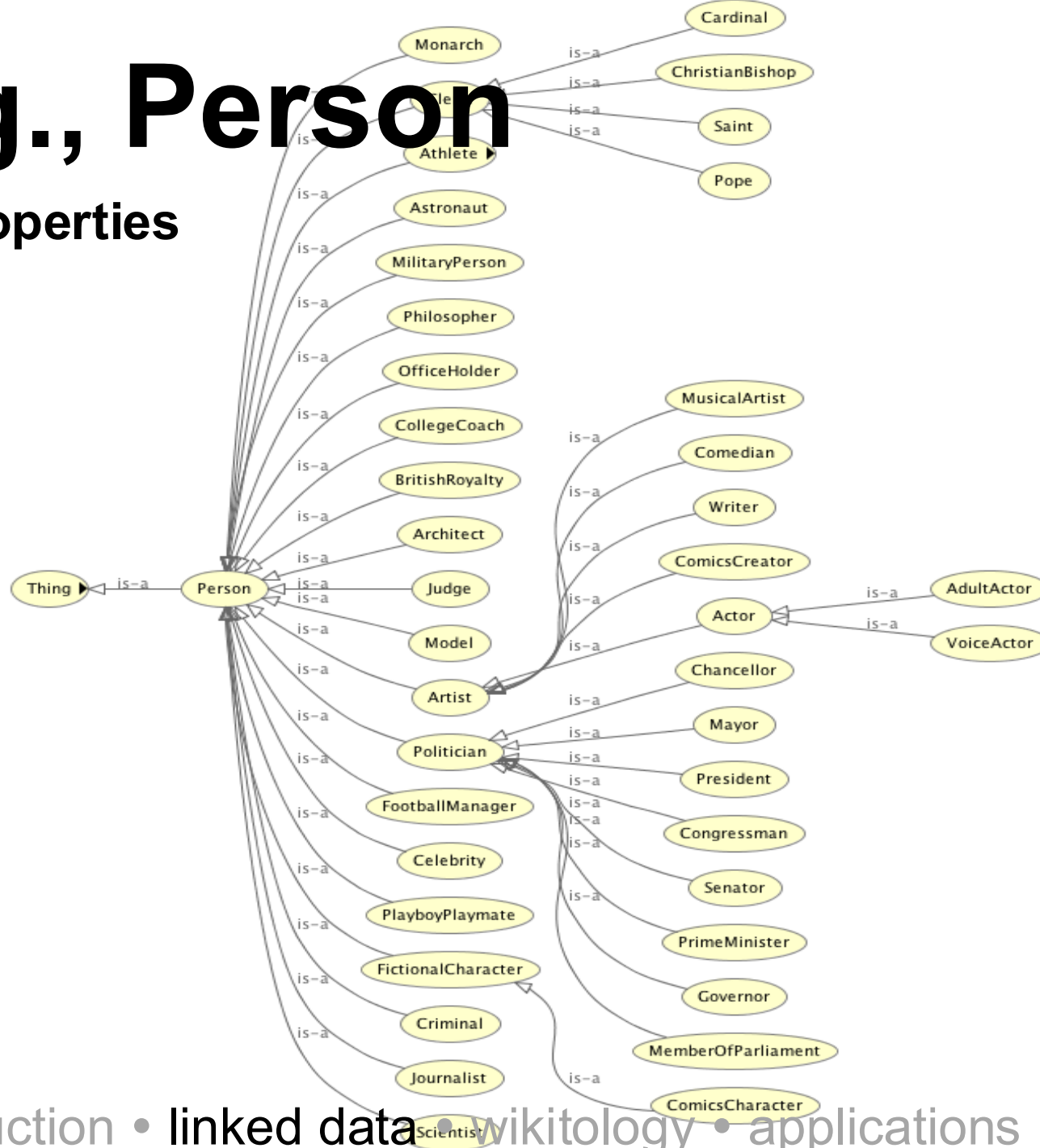
- DBpedia's representation makes the schema explicit and accessible
 - But initially inherited most of the problems in the underlying implicit schema
- Integration with the Yago ontology added richness
- Since version 3.2 (11/08) DBpedia began developing an explicit OWL ontology and mapping it to the native Wikipedia terms

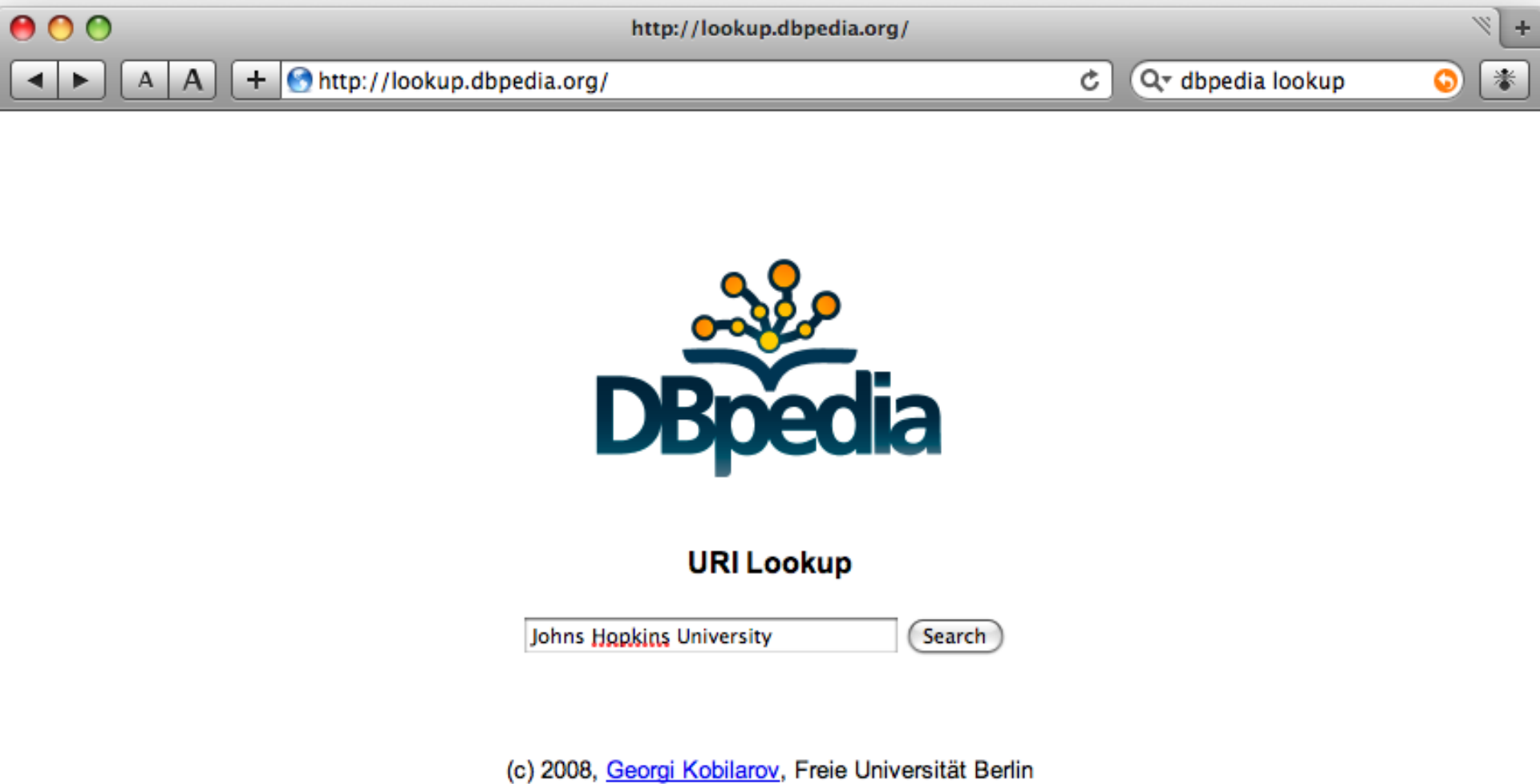
DBpedia ontology

Place	248,000
Person	214,000
Work	193,000
Species	90,000
Org.	76,000
Building	23,000

e.g., Person

56 properties





<http://lookup.dbpedia.org/>

introduction • linked data • wikitology • applications • conclusion

Johns Hopkins University EducationalInstitution Organisation Resource University

The Johns Hopkins University is a private university based in Baltimore, Maryland, United States. Johns Hopkins also maintains full-time campuses elsewhere in Maryland, Washington, D.C., Italy, and China.

Johns Hopkins University Press

The Johns Hopkins University Press is a publishing house and division of Johns Hopkins University that engages in publishing journals and books. It was founded in 1878 and holds the distinction of being the oldest continuously running university press in the United States. Its headquarters are in the Charles Village neighborhood of Baltimore, Maryland. Among the authors it has published, the following are especially noteworthy: Newt Gingrich, Beverly Whipple, Frans de Waal, Jacques Derrida, E.L. Doctorow, Donald Kennedy, Brian Lamb, Nancy Mace, H.L. Mencken, Albert Schweitzer, and E.O. Wilson.

Johns Hopkins University School of Medicine EducationalInstitution Organisation Resource University

The Johns Hopkins University School of Medicine, located in Baltimore, Maryland, U.S., is the academic medical teaching and research arm of Johns Hopkins University. The School of Medicine is widely regarded as one of the best medical schools and biomedical research institutes in the world. Located in East Baltimore, it is affiliated with Johns Hopkins Hospital, its major teaching hospital, as well as several other community sites, including the Johns Hopkins Bayview Medical Center, Sinai Hospital, Howard County General Hospital and Greater Baltimore Medical Center.

Applied Physics Laboratory

The Johns Hopkins University Applied Physics Laboratory (APL), located in Laurel, Maryland, is a not-for-profit, university-affiliated research center employing 4,150 people. APL is primarily a defense contractor.

Paul H. Nitze School of Advanced International Studies

Mission To provide a professional education that simultaneously adheres to the highest standards of scholarship and takes a practical approach to training students for international leadership. To conduct scholarly research related to the concerns of public and private institutions of the United States and governments of other countries and disseminate that research to a broad audience concerned with foreign relations. To offer mid-career educational opportunities for those already working in international affairs.

About: [Johns Hopkins University](#)



An Entity in Data Space: dbpedia.org

The Johns Hopkins University is a private university based in Baltimore, Maryland, United States. Johns Hopkins also maintains full-time campuses elsewhere in Maryland, Washington, D.C., Italy, and China.

Property	Value
dbpedia-owl:athletics	<ul style="list-style-type: none"> dbpedia:Centennial_Conference dbpedia:Division_I dbpedia:Division_III dbpedia:National_Collegiate_Athletic_Association
dbpedia-owl:city	<ul style="list-style-type: none"> dbpedia:Baltimore%2C_Maryland
dbpedia-owl:colours	<ul style="list-style-type: none"> Academic: Gold & Sable {{color box gold}} {{color box black}} Athletic: Columbia blue & Black {{color box #9BDDFF}} {{color box black}}
dbpedia-owl:country	<ul style="list-style-type: none"> dbpedia:United_States
dbpedia-owl:endowment	<ul style="list-style-type: none"> 2.8
dbpedia-owl:established	<ul style="list-style-type: none"> 1876
dbpedia-owl:faculty	<ul style="list-style-type: none"> 3100
dbpedia-owl:mascot	<ul style="list-style-type: none"> Blue Jay
dbpedia-owl:motto	<ul style="list-style-type: none"> Veritas vos liberabit (The truth shall make you free)
dbpedia-owl:postgrad	<ul style="list-style-type: none"> 14275
dbpedia-owl:president	<ul style="list-style-type: none"> dbpedia:William_R._Brody
dbpedia-owl:staff	<ul style="list-style-type: none"> 15000
dbpedia-owl:state	<ul style="list-style-type: none"> dbpedia:Maryland
dbpedia-owl:type	<ul style="list-style-type: none"> dbpedia:Private_university
dbpedia-owl:undergrad	<ul style="list-style-type: none"> 4478
dbpprop:abstract	<ul style="list-style-type: none"> The Johns Hopkins University is a private university based in Baltimore, Maryland, United


```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdfs:comment xml:lang="no">Johns Hopkins University er et privat,
ikke-kommersielt forskningsuniversitet med hovedcampus i Homewood, like nord for sentrum av Baltimore, Maryland i
USA.</rdfs:comment></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdf:type
rdf:resource="http://dbpedia.org/ontology/EducationalInstitution"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:colors xmlns:dbpprop="http://dbpedia.org/property/"
rdf:resource="http://dbpedia.org/resource/Black"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:imageName xmlns:dbpprop="http://dbpedia.org/property/"
rdf:resource="http://upload.wikimedia.org/wikipedia/en/6/67/JHU_seal.png"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdfs:label xml:lang="zh">çŸŸ·éœŸéŸ·æ-âŸŸ-
|</rdfs:label></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:colors xmlns:dbpprop="http://dbpedia.org/property/"
rdf:resource="http://dbpedia.org/resource/Columbia_blue"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><foaf:name xmlns:foaf="http://xmlns.com/foaf/0.1/">The Johns
Hopkins University</foaf:name></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpedia-owl:mascot xmlns:dbpedia-
owl="http://dbpedia.org/ontology/">Blue Jay</dbpedia-owl:mascot></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:faculty xmlns:dbpprop="http://dbpedia.org/property/"
xml:lang="en">3,100 (full time)</dbpprop:faculty></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdfs:label xml:lang="ja">ã, ãŸŸãŸã, ãŸŸãŸã, ãŸã, ãŸã, ãŸã, ãŸã-ã, ãŸã, ãŸã-
|</rdfs:label></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdfs:comment xml:lang="es">La Universidad Johns Hopkins es una
instituciÃ³n privada situada en Baltimore, Maryland, Estados Unidos.</rdfs:comment></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpedia-owl:motto xmlns:dbpedia-
owl="http://dbpedia.org/ontology/">Veritas vos liberabit&lt;br&gt;(The truth shall make you free)</dbpedia-owl:motto></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpedia-owl:faculty xmlns:dbpedia-
owl="http://dbpedia.org/ontology/">3100</dbpedia-owl:faculty></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:logo xmlns:dbpprop="http://dbpedia.org/property/"
rdf:resource="http://upload.wikimedia.org/wikipedia/en/3/38/JHUlogo.png"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:relatedInstance
xmlns:dbpprop="http://dbpedia.org/property/" rdf:resource="http://dbpedia.org/resource/Johns_Hopkins_University/convert1"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdfs:label xml:lang="fr">UniversitÃ© Johns-
Hopkins</rdfs:label></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:athletics xmlns:dbpprop="http://dbpedia.org/property/"
rdf:resource="http://dbpedia.org/resource/National_Collegiate_Athletic_Association"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:country xmlns:dbpprop="http://dbpedia.org/property/"
rdf:resource="http://dbpedia.org/resource/United_States"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdfs:comment xml:lang="pl">Uniwersytet Johnsa Hopkinsa (ang.
Johns Hopkins University) - prywatny uniwersytet w Stanach Zjednoczonych, dziaÅ,ajÅ...cy w Baltimore, w stanie
Maryland.</rdfs:comment></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpedia-owl:established xmlns:dbpedia-
owl="http://dbpedia.org/ontology/">1876</dbpedia-owl:established></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><rdfs:comment xml:lang="de">Die Johns Hopkins University (JHU) ist
eine private UniversitÃ¼t in Baltimore, USA. Sie wurde am 22.</rdfs:comment></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:relatedInstance
xmlns:dbpprop="http://dbpedia.org/property/" rdf:resource="http://dbpedia.org/resource/Johns_Hopkins_University/convert2"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:established xmlns:dbpprop="http://dbpedia.org/property/"
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">1876</dbpprop:established></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:abstract xmlns:dbpprop="http://dbpedia.org/property/"
xml:lang="nl">Johns Hopkins University is een hogere privÃ©onderwijsinstelling in Baltimore, Maryland, Verenigde Staten. Ze werd gesticht in 1876.
Johns Hopkins wordt beschouwd als Ã©Ã©n van de belangrijkste onderzoek- en onderwijsinstellingen wereldwijd. Ze scoort ook altijd hoog in
internationale rankings van prestigieuze universiteiten.</dbpprop:abstract></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns_Hopkins_University"><dbpprop:wikiPageUsesTemplate
xmlns:dbpprop="http://dbpedia.org/property/" rdf:resource="http://dbpedia.org/resource/Template:infobox_university"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Johns Hopkins University"><dbpprop:type xmlns:dbpprop="http://dbpedia.org/property/"
```

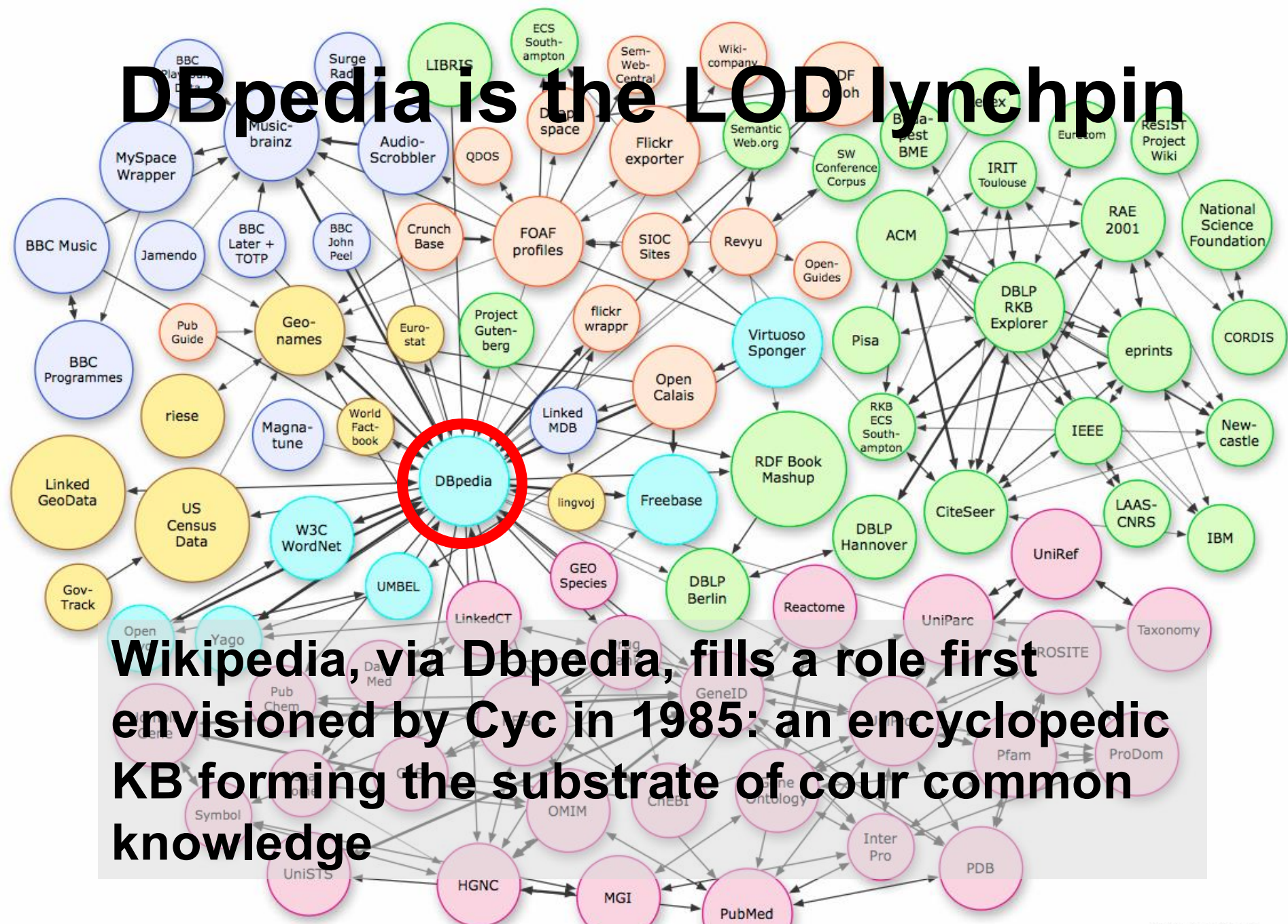
```

PREFIX dbp: <http://dbpedia.org/resource/>
PREFIX dbpo: <http://dbpedia.org/ontology/>
SELECT distinct ?Property ?Place
WHERE {dbp:Barack_Obama ?Property ?Place .
      ?Place rdf:type dbpo:Place }
    
```

Property	Place
http://dbpedia.org/property/birthPlace	http://dbpedia.org/resource/Hawaii
http://dbpedia.org/property/birthPlace	http://dbpedia.org/resource/Honolulu%2C_Hawaii
http://dbpedia.org/property/birthPlace	http://dbpedia.org/resource/United_States
http://dbpedia.org/property/state	http://dbpedia.org/resource/Illinois
http://dbpedia.org/property/nationality	http://dbpedia.org/resource/United_States
http://dbpedia.org/ontology/nationality	http://dbpedia.org/resource/United_States
http://dbpedia.org/ontology/birthplace	http://dbpedia.org/resource/Hawaii
http://dbpedia.org/ontology/birthplace	http://dbpedia.org/resource/Honolulu%2C_Hawaii
http://dbpedia.org/ontology/birthplace	http://dbpedia.org/resource/United_States

What are Barack Obama's properties with values that are places?

DBpedia is the LOD lynchpin



As of July 2009


Consider Baltimore, MD

About: Baltimore, Maryland

← → A A + http://dbpedia.org/page/Baltimore%2C_Maryland ↻ 🔍 Google

About: [Baltimore, Maryland](#)

An Entity in Data Space: dbpedia.org



Baltimore is an independent city and the largest city in the state of Maryland in the United States. Baltimore is located in central Maryland at the head of the tidal portion of the Patapsco River, an arm of the Chesapeake Bay, and is approximately 40 miles northeast of Washington D.C.. Founded in 1729, Baltimore is a major U.S. seaport and is situated closer to major Midwestern markets than any other major seaport on the East Coast.

Property	Value
dbpedia-owl:areaCode	▪ 410, 443
dbpedia-owl:areaLand	▪ 80.8
dbpedia-owl:areaTotal	▪ 92.1
dbpedia-owl:areaWater	▪ 11.3
dbpedia-owl:elevation	▪ 33
	▪ 10
dbpedia-owl:establishedTitle	▪ Founded
dbpedia-owl:foundingDate	▪ 1729
dbpedia-owl:latitudeminutes	▪ 17
dbpedia-owl:latitudenorthorsouth	▪ N
dbpedia-owl:latitudeseconds	▪ 11
dbpedia-owl:latitutedegrees	▪ 39
dbpedia-owl:leaderName	▪ dbpedia:Sheila_Dixon
	▪ dbpedia:United_States_Democratic_Party
dbpedia-owl:leaderTitle	▪ Mayor
dbpedia-owl:longitudedegrees	▪ 76
dbpedia-owl:longitudeminutes	▪ 36
dbpedia-owl:longitudeseastorwest	▪ W
dbpedia-owl:longitudeseconds	▪ 54
dbpedia-owl:motto	▪ "The Greatest City in America", "Get in on it." "Believe." (formerly "The City That Reads")
dbpedia-owl:nickname	▪ Charm City, Mob Town, Crabtown, B-more, The City of Firsts, Monument City, B-Town, Ravenstown
dbpedia-owl:otherName	▪ Charm City, Mob Town, Crabtown, B-more, The City of Firsts, Monument City, B-Town, Ravenstown
dbpedia-owl:populationAsOf	▪ 2007
dbpedia-owl:populationDensity	▪ 8058.4
dbpedia-owl:populationMetro	▪ 2668056
dbpedia-owl:populationTotal	▪ 637455
dbpedia-owl:populationUrbanTotal	▪ 2178000

Links between RDF datasets

- We find assertions equating DBpedia's Baltimore object with those in other LOD datasets

```
dbpedia:Baltimore%2C_Maryland
```

```
owl:sameAs census:us/md/counties/baltimore/baltimore;
```

```
owl:sameAs cyc:concept/Mx4rvVin-5wpEbGdrcN5Y29ycA;
```

```
owl:sameAs freebase:guid.9202a8c04000641f8000004921a;
```

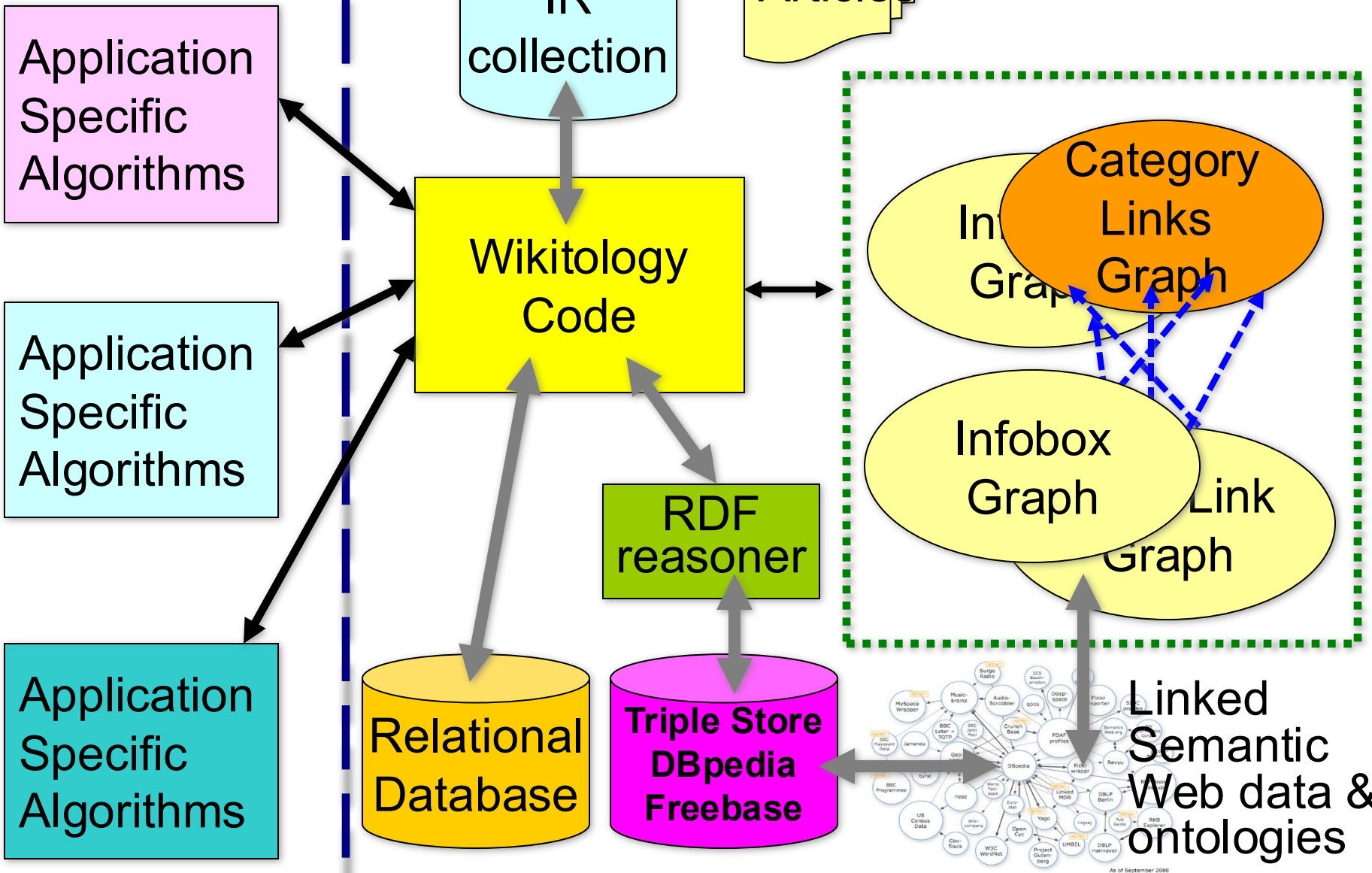
```
owl:sameAs geonames:4347778/ .
```

- Since owl:sameAs is defined as an equivalence relation, the mapping works both ways
- Mappings are done by custom programs, machine learning, and manual techniques

Wikitology

- We've explored a complementary approach to derive an ontology from Wikipedia: Wikitology
- Wikitology use cases:
 - Identifying user context in a collaboration system from documents viewed (2006)
 - Improve IR accuracy of by adding Wikitology tags to documents (2007)
 - ACE: cross document co-reference resolution for named entities in text (2008)
 - TAC KBP: Knowledge Base population from text (2009)

Wikitology 3.0 (2009)



As of September 2008

Wikitology

- We've explored a complementary approach to derive an ontology from Wikipedia: Wikitology
- Wikitology use cases:
 - Identifying user context in a collaboration system from documents viewed (2006)
 - Improve IR accuracy of by adding Wikitology tags to documents (2007)
 - ACE 2008: cross document co-reference resolution for named entities in text (2008)
 - TAC 2009: Knowledge Base population from text (2009)

ACE 2008: Cross-Document Coreference Resolution

- Determine when two documents mention the same entity
 - Are two documents that talk about “George Bush” talking about the same George Bush?
 - Is a document mentioning “Mahmoud Abbas” referring to the same person as one mentioning “Muhammed Abbas”? What about “Abu Abbas”? “Abu Mazen”?
- Drawing appropriate inferences from multiple documents demands *cross-document coreference resolution*



ACE 2008: Wikitology tagging

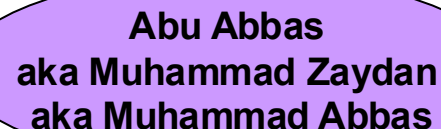
- NIST ACE 2008: cluster named entity mentions in 20K English and Arabic documents
- We produced an *entity document* for mentions with name, nominal and pronominal mentions, type and subtype, and nearby words
- Tagged these with Wikitology producing vectors to compute features measuring entity pair similarity
- One of many features for an SVM classifier



William Wallace
(living British Lord)



William Wallace
(of Braveheart fame)



Abu Abbas
aka Muhammad Zaydan
aka Muhammad Abbas

Wikitology Entity Document & Tags

Wikitology entity document

<DOC>
<DOCNO>ABC19980430.1830.0091.LDC2000T44-E2 <DOCNO>
<TEXT>
Webb Hubbell ← **Name**
PER ← **Type & subtype**
Individual
NAM: "Hubbell" "Hubbells" "Webb Hubbell" "Webb_Hubbell"
PRO: "he" "him" "his" ← **Mention heads**
abc's accountant after again ago all alleges alone also and arranged
attorney avoid been before being betray but came can cat charges
cheating circle clearly close concluded conspiracy cooperate counsel
counsel's department did disgrace do dog dollars earned eightynine
enough evasion feel financial firm first four friend friends going got
grand happening has he help him hi s hope house hubbell hubbells
hundred hush income increase independent indict indicted indictment
inner investigating jackie jackie_judd jail jordan judd jury justice
kantor ken knew lady late law left lie little make many mickey mid
money mr my nineteen nineties ninetyfour not nothing now office
other others paying peter_jennings president's pressure pressured
probe prosecutors questions reported reveal rock saddened said
schemed seen seven since starr statement such tax taxes tell them
they thousand time today ultimately vernon washington webb
webb_hubbell were what's whether which white whitewater why wife
years
</TEXT>
</DOC>

← **Words surrounding mentions**

Wikitology article tag vector

Webster_Hubbell 1.000
Hubbell_Trading_Post National Historic Site 0.379
United_States_v._Hubbell 0.377
Hubbell_Center 0.226
Whitewater_controversy 0.222

Wikitology category tag vector

Clinton_administration_controversies 0.204
American_political_scandals 0.204
Living_people 0.201
1949_births 0.167
People_from_Arkansas 0.167
Arkansas_politicians 0.167
American_tax_evaders 0.167
Arkansas_lawyers 0.167

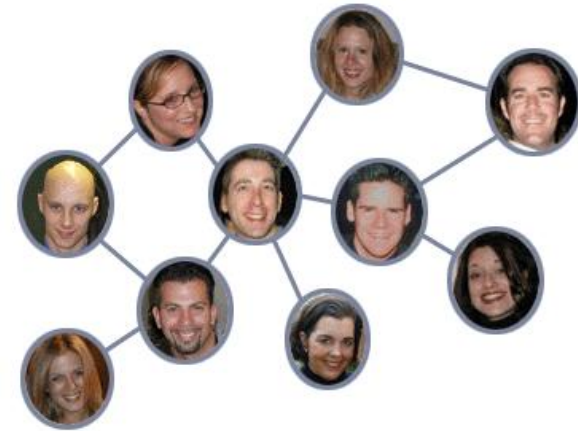
Top Ten Features (by F1)

Prec.	Recall	F1	Feature Description
90.8%	76.6%	83.1%	some NAM mention has an exact match
92.9%	71.6%	80.9%	Dice score of NAM strings (based on the intersection of NAM strings, not words or n-grams of NAM strings)
95.1%	65.0%	77.2%	the/a longest NAM mention is an exact match
86.9%	66.2%	75.1%	Similarity based on cosine similarity of Wikitology Article Medium article tag vector
86.1%	65.4%	74.3%	Similarity based on cosine similarity of Wikitology Article Long article tag vector
64.8%	82.9%	72.8%	Dice score of character bigrams from the 'longest' NAM string
95.9%	56.2%	70.9%	all NAM mentions have an exact match in the other pair
85.3%	52.5%	65.0%	Similarity based on a match of entities' top Wikitology article tag
85.3%	52.3%	64.8%	Similarity based on a match of entities' top Wikitology article tag
85.7%	32.9%	47.5%	Pair has a known alias

The Wikitology-based features were very useful

Wikipedia's Social Network

- Wikipedia has an implicit 'social network' that can help disambiguate PER mentions (ORGs & GPEs too)
- We extracted 875K people from Freebase, 616K of were linked to Wikipedia pages, 431K of which are in one of 4.8M person-person article links
- Consider a document that mentions two people: *George Bush* and *Mr. Quayle*
- There are six George Bushes in Wikipedia and nine Male Quayles



Which Bush & which Quayle?



Six George Bushes

Nine Male Quayles

Use Jaccard coefficient metric

Let $S_i = \{\text{two hop neighbors of } S_i\}$

$C_{ij} = |\text{intersection}(S_i, S_j)| / |\text{union}(S_i, S_j)|$

$C_{ij} > 0$ for six of the 56 possible pairs

0.43 George_H._W._Bush -- Dan_Quayle

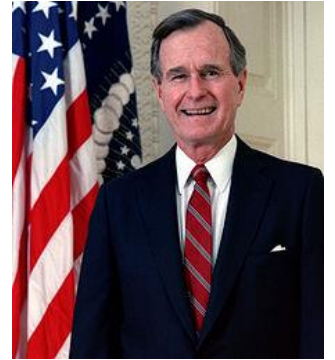
0.24 George_W._Bush -- Dan_Quayle

0.18 George_Bush_(biblical_scholar) -- Dan_Quayle

0.02 George_Bush_(biblical_scholar) -- James_C._Quayle

0.02 George_H._W._Bush -- Anthony_Quayle

0.01 George_H._W._Bush -- James_C._Quayle



Knowledge Base Population

- The 2009 NIST Text Analysis Conference had a Knowledge Base Population track
 - Add facts to a reference KB from a collection of 1.3M English newswire documents
- Given initial KB of facts from Wikipedia info-boxes: 200k people, 200k GPEs, 60k orgs, 300+k misc/non-entities
- Two fundamental tasks:
 - **Entity Linking** - Grounding entity mentions in documents to KB entries (or NIL if not in KB)
 - **Slot Filling** - Learning additional attributes

Sample KB Entry

```
<entity wiki_title="Michael_Phelps"
  type="PER"
  id="E0318992"
  name="Michael Phelps">
<facts class="Infobox Swimmer">
<fact name="swimmername">Michael Phelps</fact>
<fact name="fullname">Michael Fred Phelps</fact>
<fact name="nicknames">The Baltimore Bullet</fact>
<fact name="nationality">United States</fact>
<fact name="strokes">Butterfly, Individual Medley, Freestyle, Backstroke</fact>
<fact name="club">Club Wolverine, University of Michigan</fact>
<fact name="birthdate">June 30, 1985 (1985-06-30) (age 23)</fact>
<fact name="birthplace">Baltimore, Maryland, United States</fact>
<fact name="height">6 ft 4 in (1.93 m)</fact>
<fact name="weight">200 pounds (91 kg)</fact>
</facts>
<wiki_text><![CDATA[Michael Phelps
Michael Fred Phelps (born June 30, 1985) is an American swimmer. He has
Olympic gold medals, the most by any Olympian. As of August 2008, he also
world records in swimming. Phelps holds the record for the most gold medal
single Olympics with the eight golds he won at the 2008 Olympic Games...
```

Michael Phelps



Michael Phelps at the 2008 Beijing Olympics

Personal information

Full name:	Michael Fred Phelps
Nickname(s):	The Baltimore Bullet ^[1]
Nationality:	 United States
Stroke(s):	Butterfly, Individual Medley, Freestyle, Backstroke
Club:	Club Wolverine, University of Michigan
Date of birth:	June 30, 1985 (age 23)
Place of birth:	Baltimore, Maryland, United States
Height:	6 ft 4 in (1.93 m)
Weight:	200 pounds (91 kg)

Medal record [\[show\]](#)

Entity Linking Task



John Williams

Richard Kaufman goes a long way back with **John Williams**. Trained as a classical violinist, Californian Kaufman started doing session work in the Hollywood studios in the 1970s. One of his movies was Jaws, with **Williams** conducting his score in recording sessions in 1975...

John Williams	author	1922-1994
J. Lloyd Williams	botanist	1854-1945
John Williams	politician	1955-
John J. Williams	US Senator	1904-1988
John Williams	Archbishop	1582-1650
John Williams	composer	1932-
Jonathan Williams	poet	1929-

Michael Phelps

Debbie Phelps, the mother of swimming star **Michael Phelps**, who won a record eight gold medals in Beijing, is the author of a new memoir, ...

Michael Phelps is the scientist most often identified as the inventor of PET, a technique that permits the imaging of biological processes in the organ systems of living individuals. **Phelps** has ...

Michael Phelps	swimmer	1985-
Michael Phelps	biophysicist	1939-

Identify matching entry, or determine that entity is missing from KB

Slot Filling Task

Target: EPA
+ context document

Generic Entity Classes
Person, Organization, GPE

Environmental Protection Agency



Agency overview

Employees	17,964 (2005)
Annual budget	\$7.3 billion (2007)
Agency executive	Lisa P. Jackson, Administrator

Missing information to mine from text:

- **Date formed:** 12/2/1970
- **Website:** <http://www.epa.gov/>
- **Headquarters:** Washington, DC
- **Nicknames:** EPA, USEPA
- **Type:** federal agency
- **Address:** 1200 Pennsylvania Avenue NW

Optional: Link some learned values within the KB:

- **Headquarters:** Washington, DC (kbid: 735)

KB Entity Attributes

Person	Organization	Geo-Political Entity
alternate names	alternate names	alternate names
age	political/religious affiliation	capital
birth: date, place	top members/employees	subsidiary orgs
death: date, place, cause	number of employees	top employees
national origin	members	political parties
residences	member of	established
spouse	subsidiaries	population
children	parents	currency
parents	founded by	
siblings	founded	
other family	dissolved	
schools attended	headquarters	
job title	shareholders	
employee-of	website	
member-of		
religion		
criminal charges		

HLTCOE* Entity Linking: Approach

** Human Language Technology Center of Excellence*

- Two-phased approach
 1. Candidate Set Identification
 2. Candidate Ranking
- Candidate Set Identification
 - Small set of easy-to-compute features
 - Speed linear in size of KB (~700K entities)
 - Constant-time possible, though recall could fall
- Candidate Ranking
 - Supervised machine learning (SVM)
 - Goal is to rank candidates
 - ~~Many features~~ Many, many features
 - Experimental development with 100s tests on held-out data

Phase 1: Candidate Identification



- 'Triage' features:
 - String comparison
 - Exact/Fuzzy String match, Acronym match
 - Known aliases
 - Wikipedia redirects provide rich set of alternate names



- Statistics
 - 98.6% recall (vs. 98.8% on dev. data)
 - Median = 15 candidates; Mean = 76; Max = 2772
 - 10% of queries ≤ 4 candidates; 10% > 100 candidates
 - Four orders of magnitude reduction in number of entities considered



Candidate Phase Failures

- Iron Lady
 - EL 1687: refers to Yulia Tymoshenko (prime minister)
 - EL 1694: refers to Biljana Plavsic (war criminal)
- PCC
 - EL 2885: Cuban Communist Party (in Spanish: *Partido Comunista de Cuba*)
- Queen City
 - EL 2973: Manchester, NH (active nickname)
 - EL 2974: Seattle, WA (former nickname)
- The Lions
 - EL 3402: Highveld Lions (South African professional cricket team) in KB as: 'Highveld_Lions_cricket_team'



Phase 2: Candidate Ranking

- Supervised Machine Learning
 - SVMrank (Joachims)
 - Trained on 1615 examples
 - About 200 atomic features, most binary
 - Cost function:
 - Number of swaps to elevate correct candidate to top of ranked list
 - “None of the above” (NIL) is an acceptable choice

“According to the CDC the prevalence of H1N1 influenza in California prisons has...”

“William C. Norris, 95, founder of the mainframe computer firm CDC., died Aug. 21 in a nursing home ... ”

Query = “CDC”

1. California Dept. of Corrections
2. US Center for Disease Control
3. Cedar City Regional Airport (IATA code)
4. Communicable Disease Centre (Singapore)
5. Congress for Democratic Change (Liberian political party)
6. Cult of the Dead Cow (Hacker organization)
7. Control Data Corporation
8. NIL (Absence from KB)
9. Consumers for Dental Choice (non-profit)
10. Cheerdance Competition (Philippine organization)

Results: top five systems

Team	All	in KB	NIL	
Siel_093	0.8217	0.7654	0.8641	Int. Inst. Of IT, Hyderabad IN
QUANTA1	0.8033	0.7725	0.8264	Tsinghua University
hltcoe1	0.7984	0.7063	0.8677	
Stanford_UBC2	0.7884	0.7588	0.8107	
NLPR_KBP1	0.7672	0.6925	0.8232	Institute for PR, China
'NIL' Baseline	0.5710	0.0000	1.0000	

Micro-averaged accuracy

Of the 13 entrants, the HLTCOE system placed third, but the differences between 2, 3 and 4 are not significant

KBP Conclusions

- Significant reductions in number of KB nodes examined possible with minimal loss of recall
- Supervised machine learning with a variety of features over query/KB node pairs is effective
- More features is better; Wikitology features were largely redundant with KB
- Optimal feature set selection varies with likelihood that query targets are in KB

Conclusions

- The Web has made people *smarter* and more *capable*, providing easy access to the world's knowledge and services
- Software agents need better access to a Web of data and knowledge to enhance their intelligence
- Some key technologies are ready to exploit: Semantic Web, linked data, RDF search engines, DBpedia, Wikitology, information extraction, etc.

Conclusion

- Hybrid systems like Wikitology combining IR, RDF, and custom graph algorithms are promising
- The linked open data (LOD) collection is a good source of *background knowledge*, useful in many tasks, e.g., extracting information from text
- The techniques can support distributed LOD collections for your domain: bioinformatics, finance, eco-informatics, etc.

<http://ebiquity.umbc.edu/>