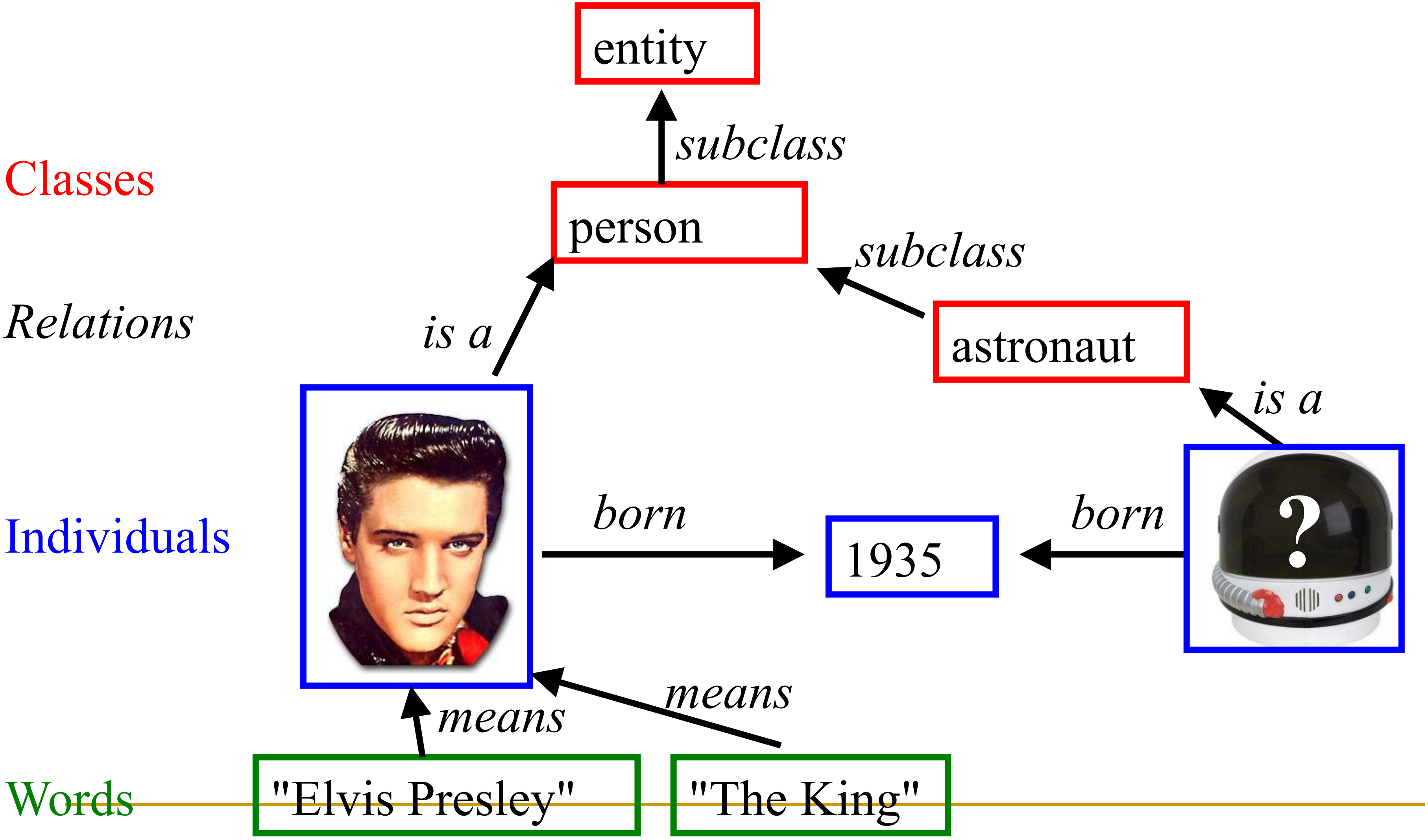
Mining Meaning From Wikipedia:

Yago – DBPedia - EMLR

PD Dr. Günter Neumann

LT-lab, DFKI, Saarbrücken

Solution: An ontology



Where do we get the ontology from?

Previous approaches:

- Assemble the ontology manually

  (WordNet, SUMO, GeneOntology)

  Problems: Usually low coverage (MPI is in none of these)

- Extract the ontology from corpora (e.g. the Web)

  (KnowItAll, Espresso, Snowball, LEILA)

  Problem: Usually low accuracy (50%-92%)

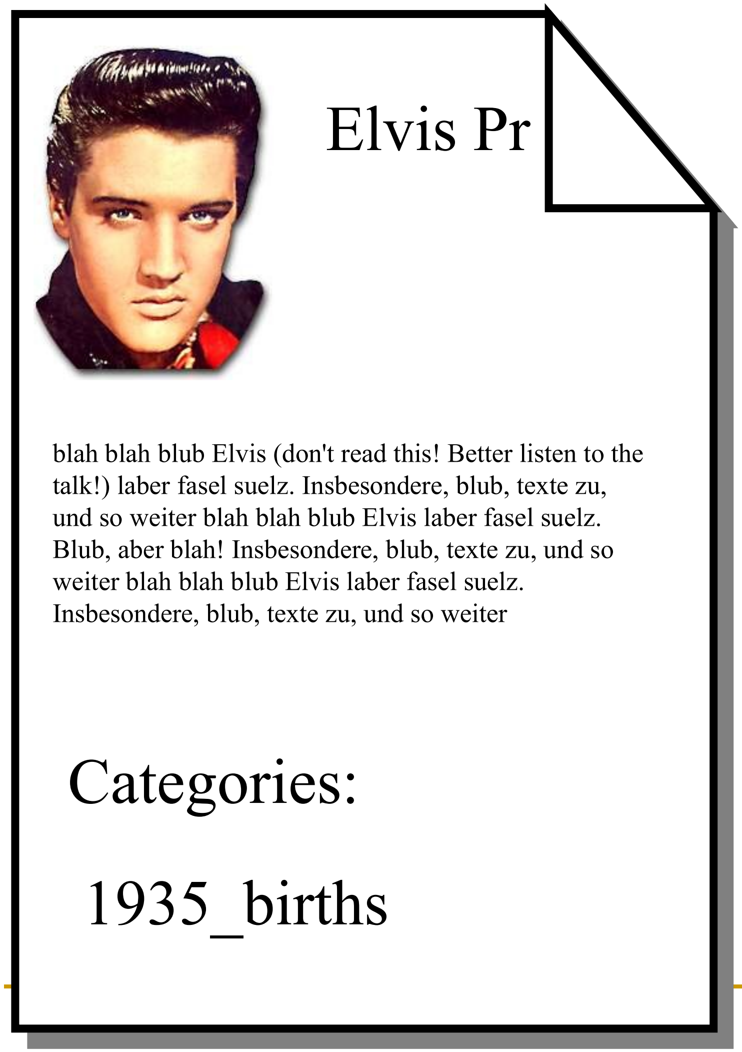
Where do we get the ontology from?

YAGO approach:

Assemble the ontology from Wikipedia (=> good coverage)

Use the category system of Wikipedia (=> good accuracy)

Exploiting the Wikipedia category system

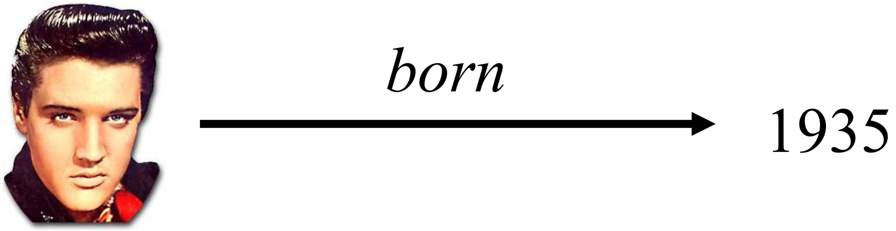


Elvis Pr

blah blah blub Elvis (don't read this! Better listen to the talk!) laber fasel suelz. Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Blub, aber blah! Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Insbesondere, blub, texte zu, und so weiter

Categories:

1935_births

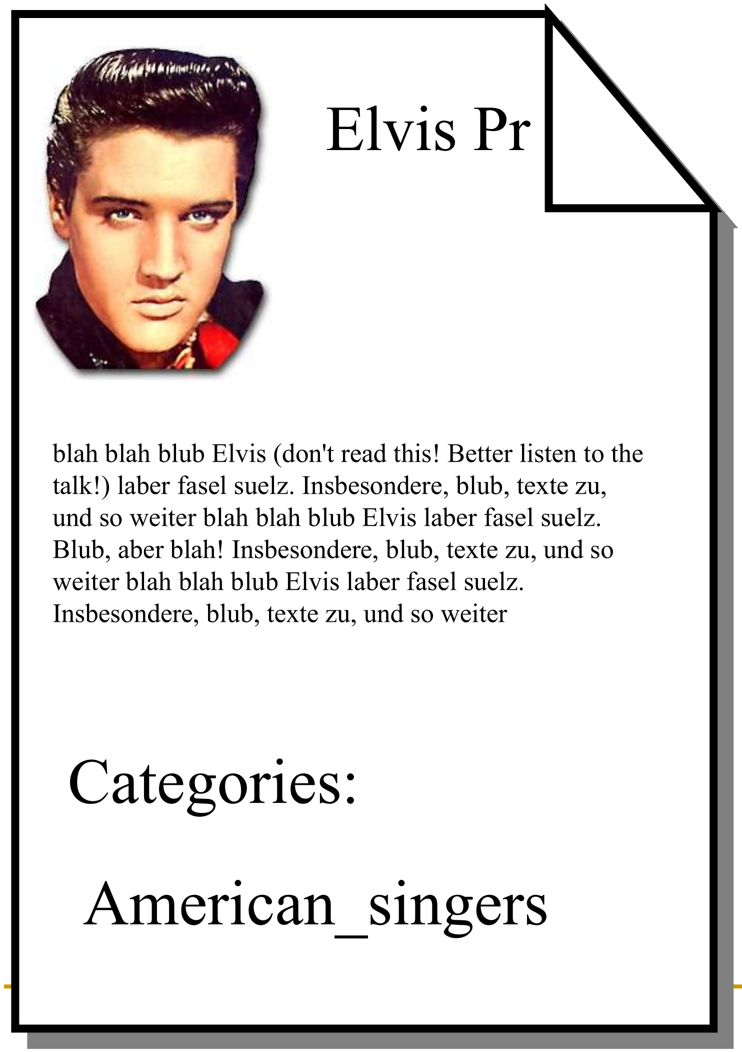


born

1935

Exploit relational categories

Exploiting the Wikipedia category system

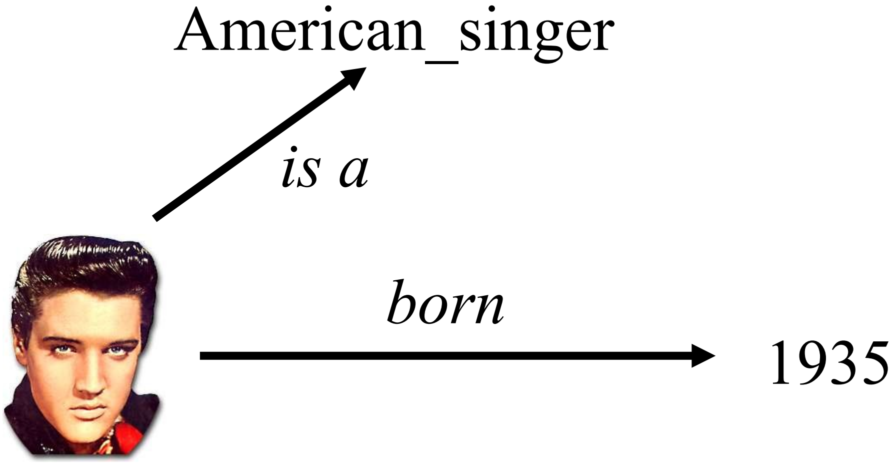


Elvis Pr

blah blah blub Elvis (don't read this! Better listen to the talk!) laber fasel suelz. Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Blub, aber blah! Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Insbesondere, blub, texte zu, und so weiter

Categories:

American_singers



American_singer

is a

born

1935

Exploit relational categories

Exploit conceptual categories

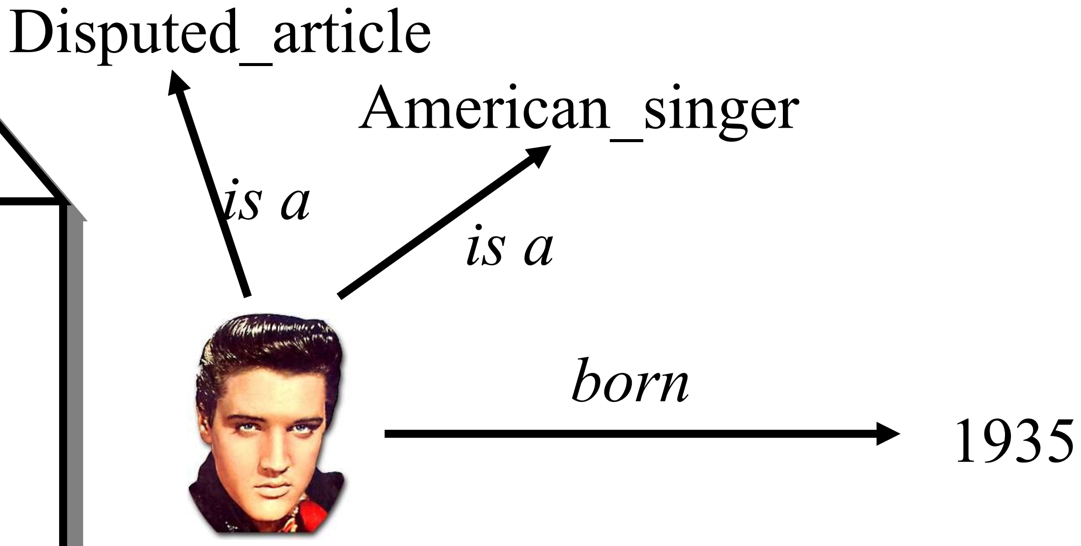
Exploiting the Wikipedia category system

Elvis Pr

blah blah blub Elvis (don't read this! Better listen to the talk!) laber fasel suelz. Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Blub, aber blah! Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Insbesondere, blub, texte zu, und so weiter

Categories:

Disputed_articles



Disputed_article

American_singer

is a

is a

born

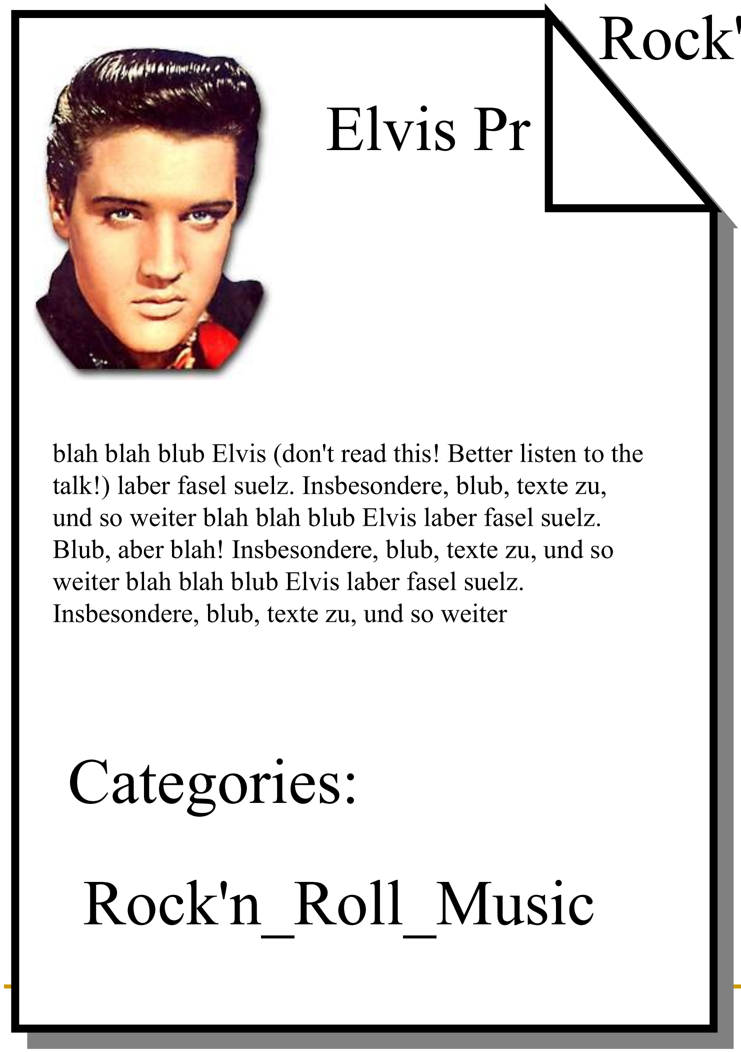
1935

Exploit relational categories

Exploit conceptual categories

Avoid administrational categories

Exploiting the Wikipedia category system

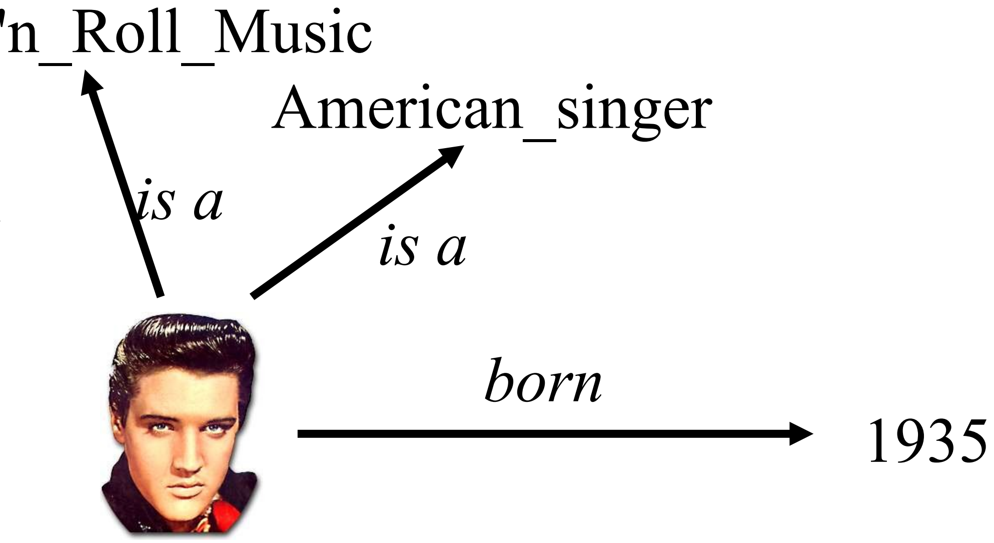


Elvis Pr

blah blah blub Elvis (don't read this! Better listen to the talk!) laber fasel suelz. Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Blub, aber blah! Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Insbesondere, blub, texte zu, und so weiter

Categories:

Rock'n_Roll_Music



Exploit relational categories

Exploit conceptual categories

Avoid administrational categories

Avoid thematic categories

Thematic vs Conceptual Categories

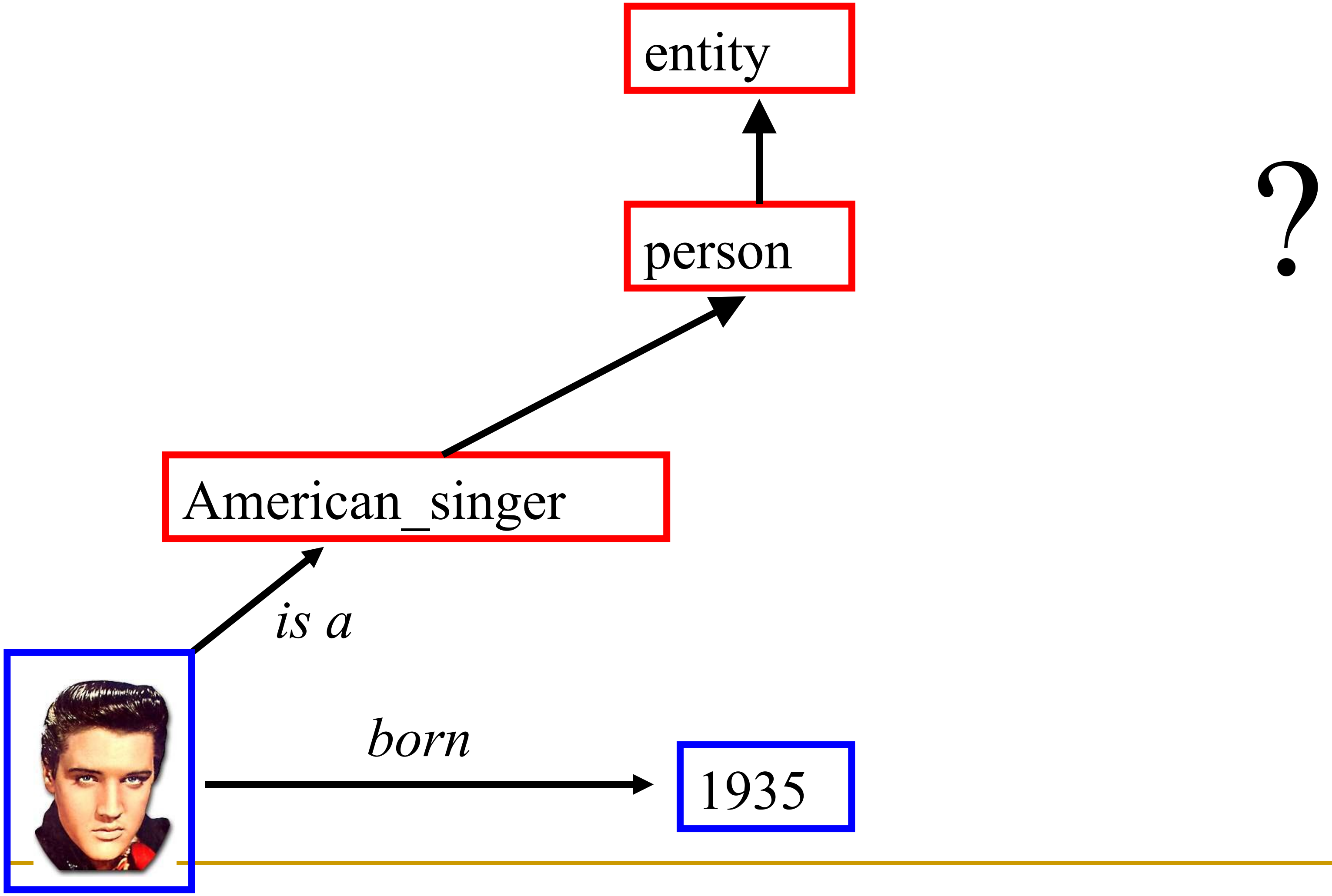
American singers | of | German origin

Shallow linguistic noun phrase parsing:

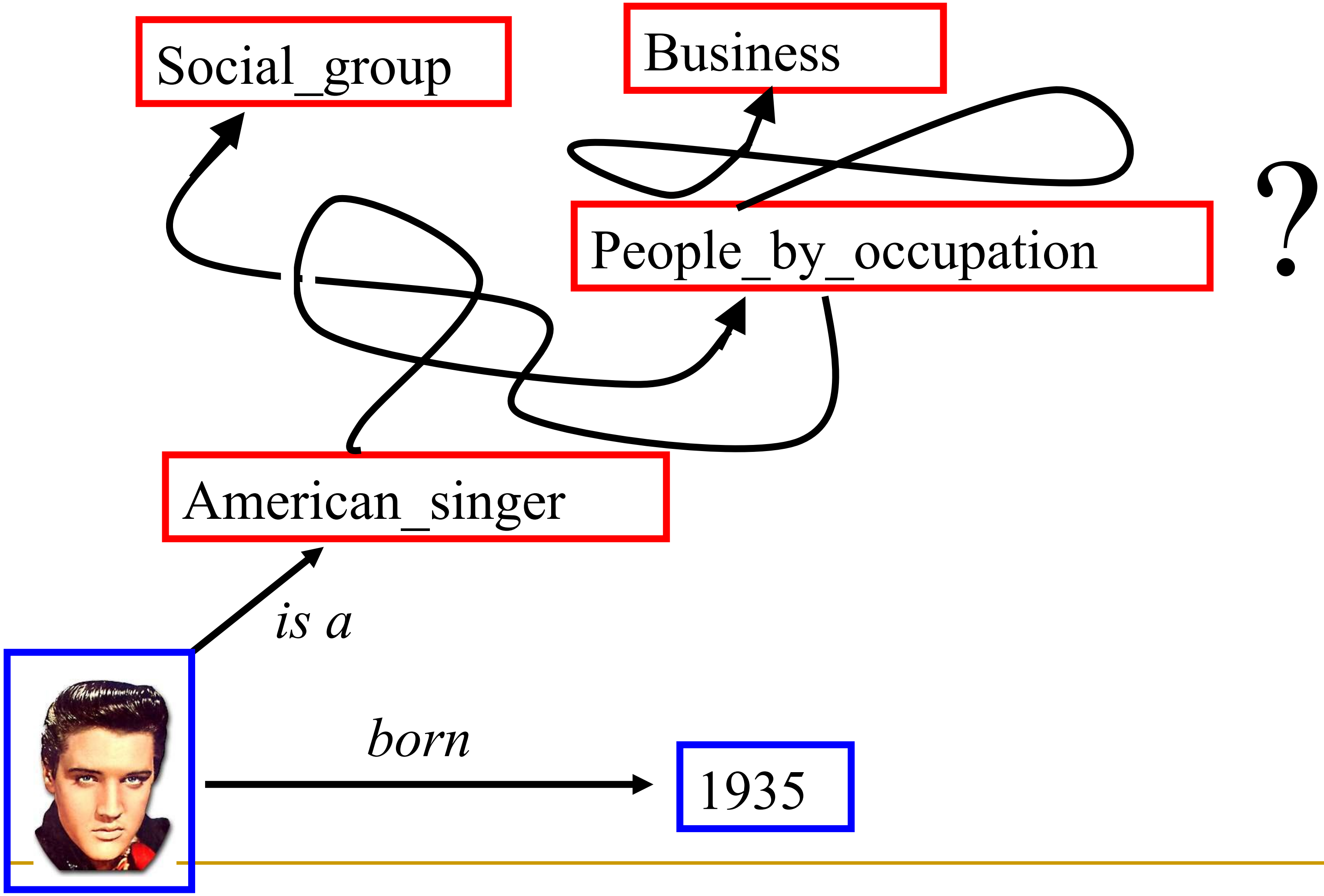
Premodifier | *Head* | *Postmodifier*

Heuristics: If the head is a plural word, the category is conceptual

The Upper Model



The Upper Model: From Wikipedia?



Social_group

Business

People_by_occupation

?

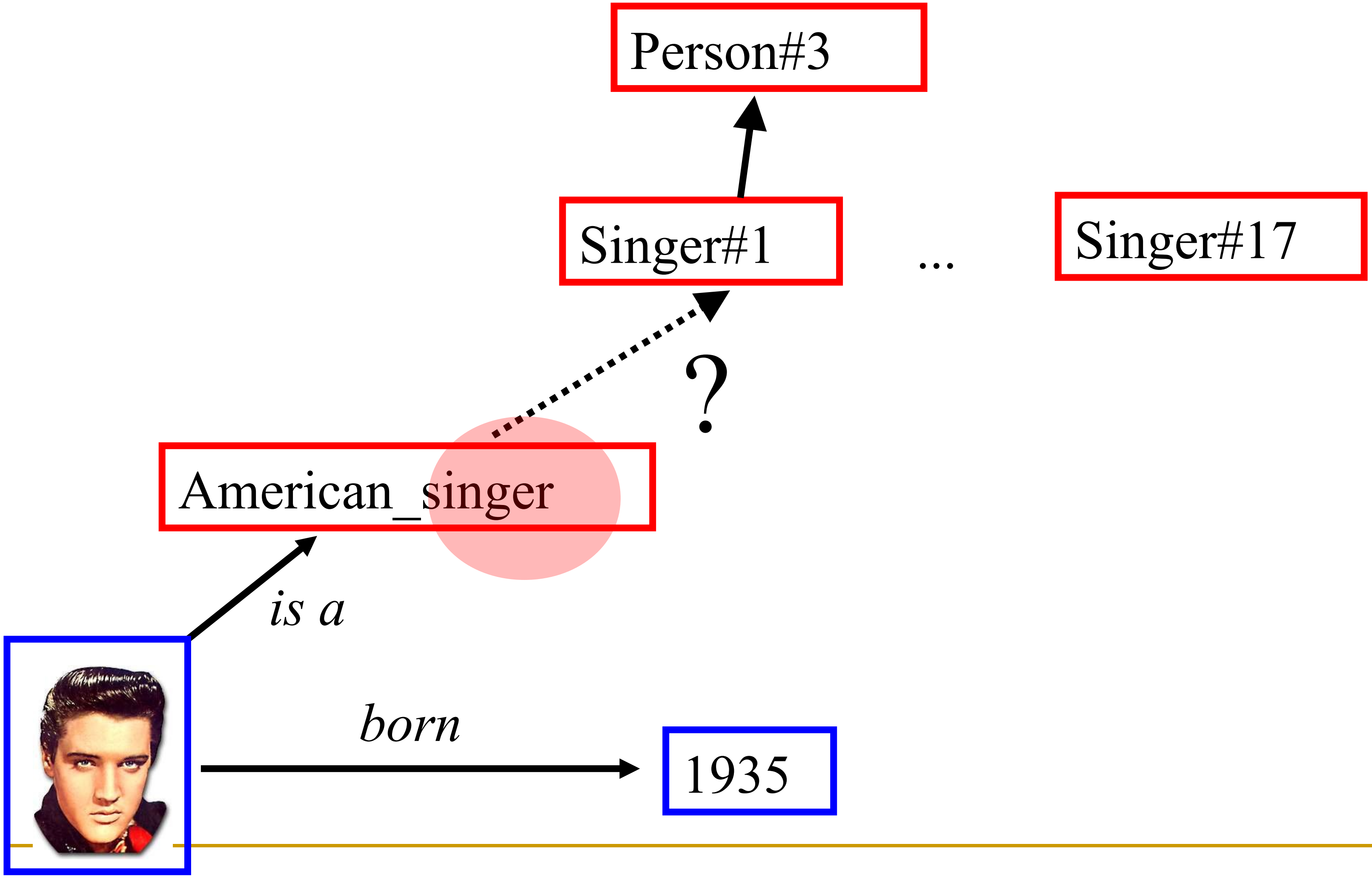
American_singer

is a

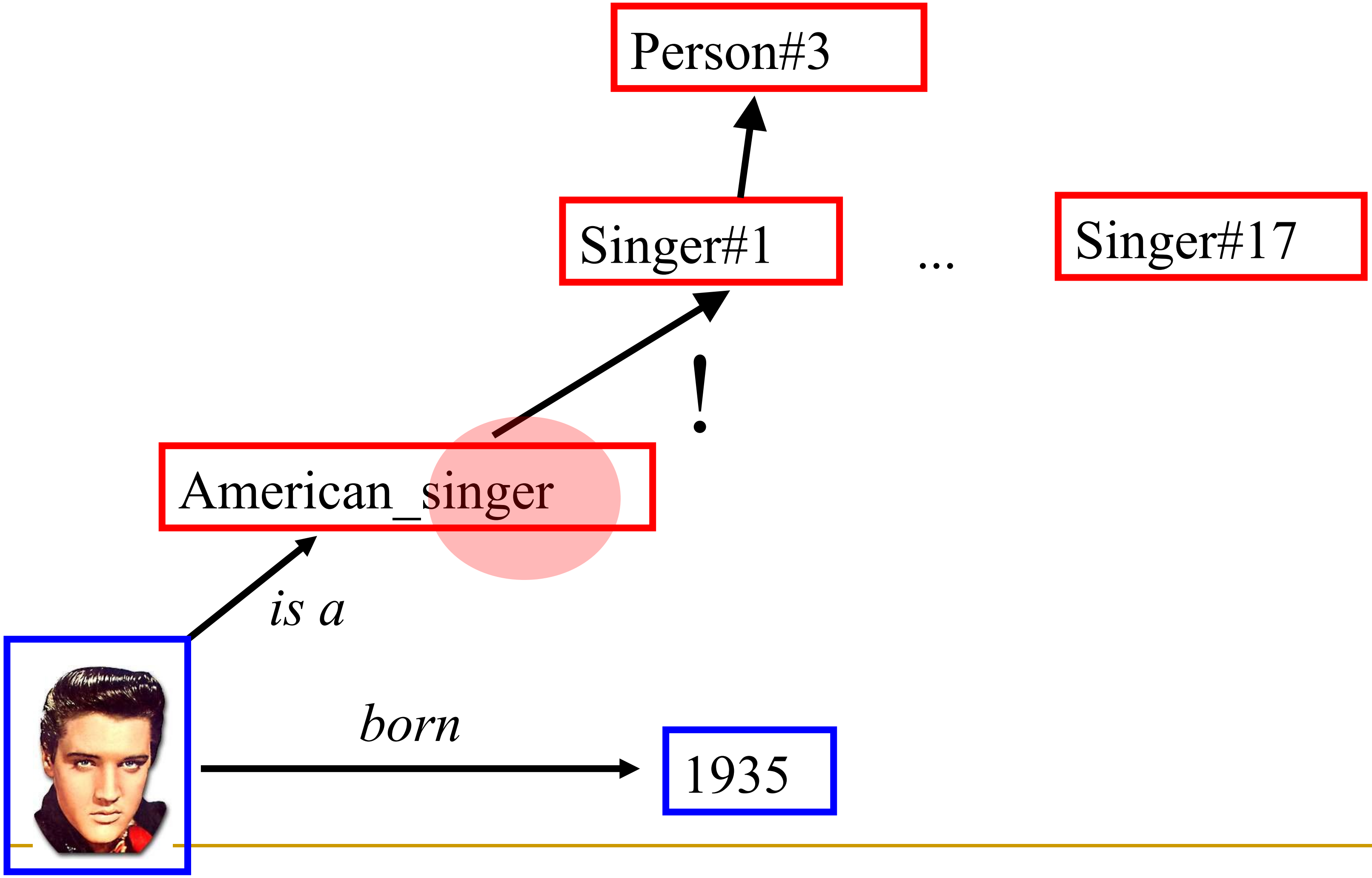
born

1935

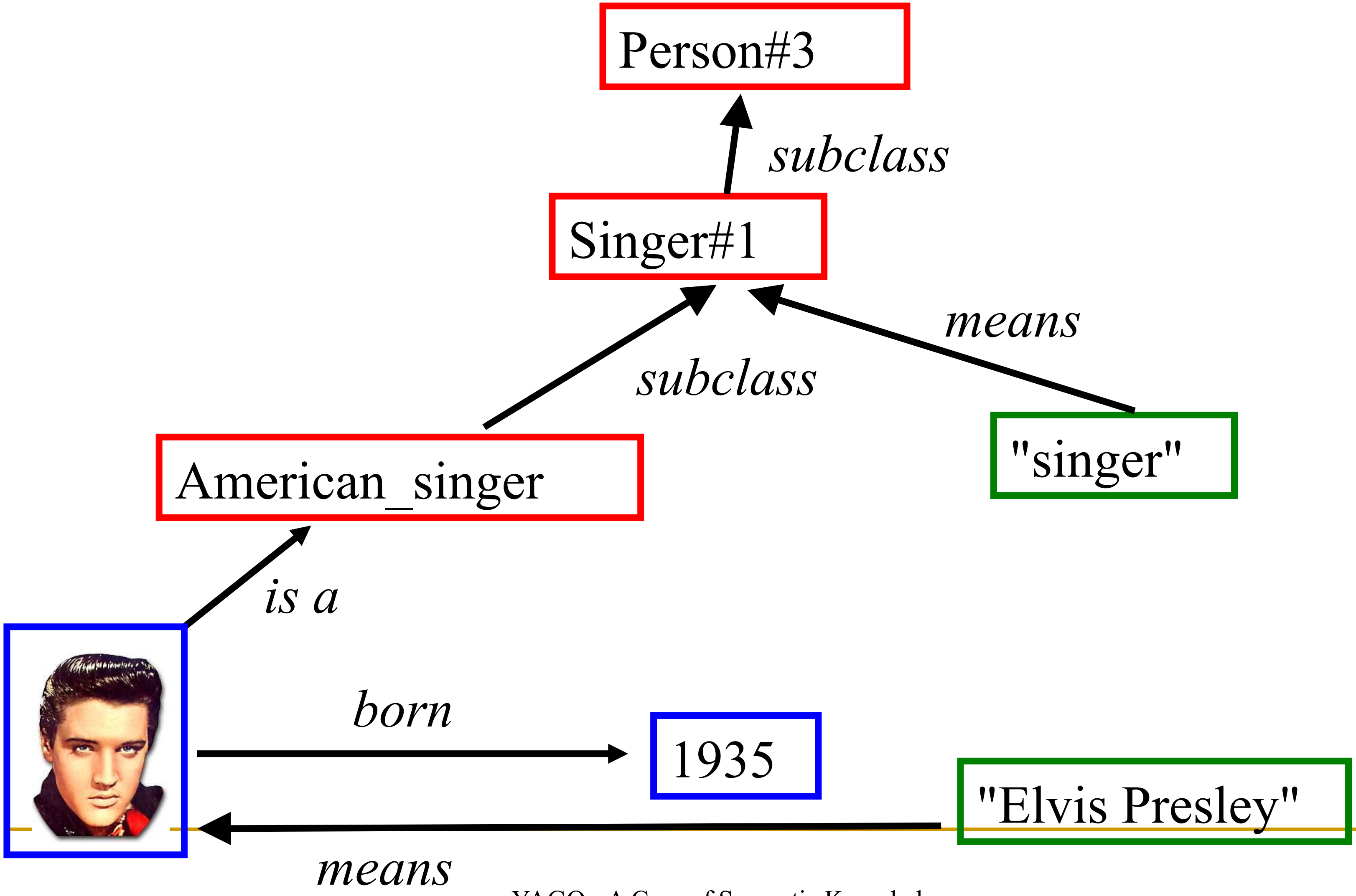
The Upper Model: From WordNet?



The Upper Model: From WordNet?



The YAGO ontology



Person#3

subclass

Singer#1

means

subclass

"singer"

American_singer

is a

born

1935

"Elvis Presley"

means

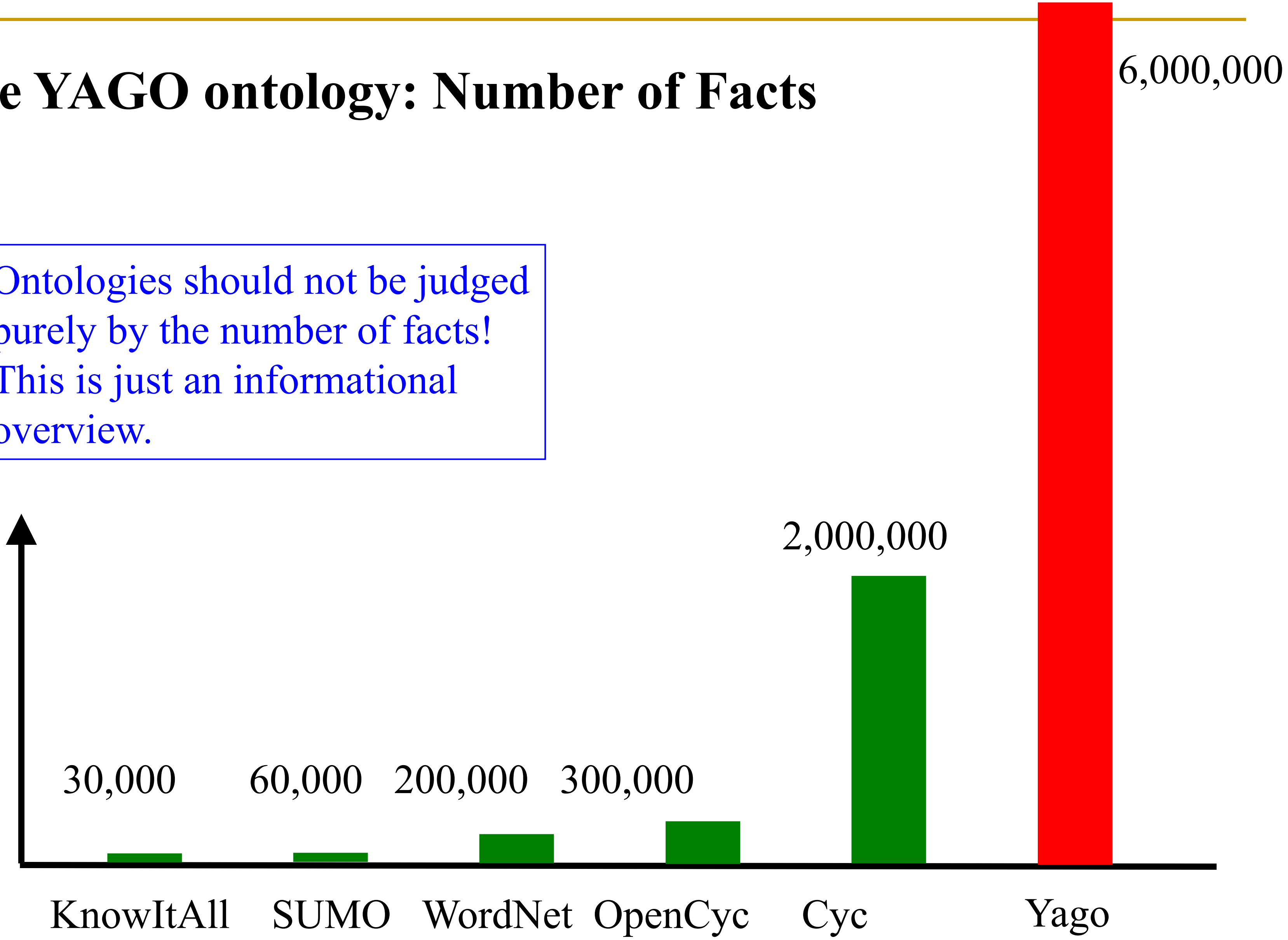
The YAGO ontology: Accuracy

Relation	Accuracy
subclass	97.70% +/- 1.59%
is a	94.54% +/- 2.36%
familyName	97.81% +/- 1.75%
givenName	97.62% +/- 2.08%
establishedIn	90.84% +/- 4.28%
bornInYear	93.14% +/- 3.71%
diedInYear	98.72% +/- 1.30%
locatedIn	98.41% +/- 1.52%
politicianOf	92.43% +/- 3.93%
writtenInYear	94.35% +/- 3.33%
hasWonPrize	98.47% +/- 1.57%

See TechReport for details on the evaluation.

The YAGO ontology: Number of Facts

Ontologies should not be judged purely by the number of facts! This is just an informational overview.



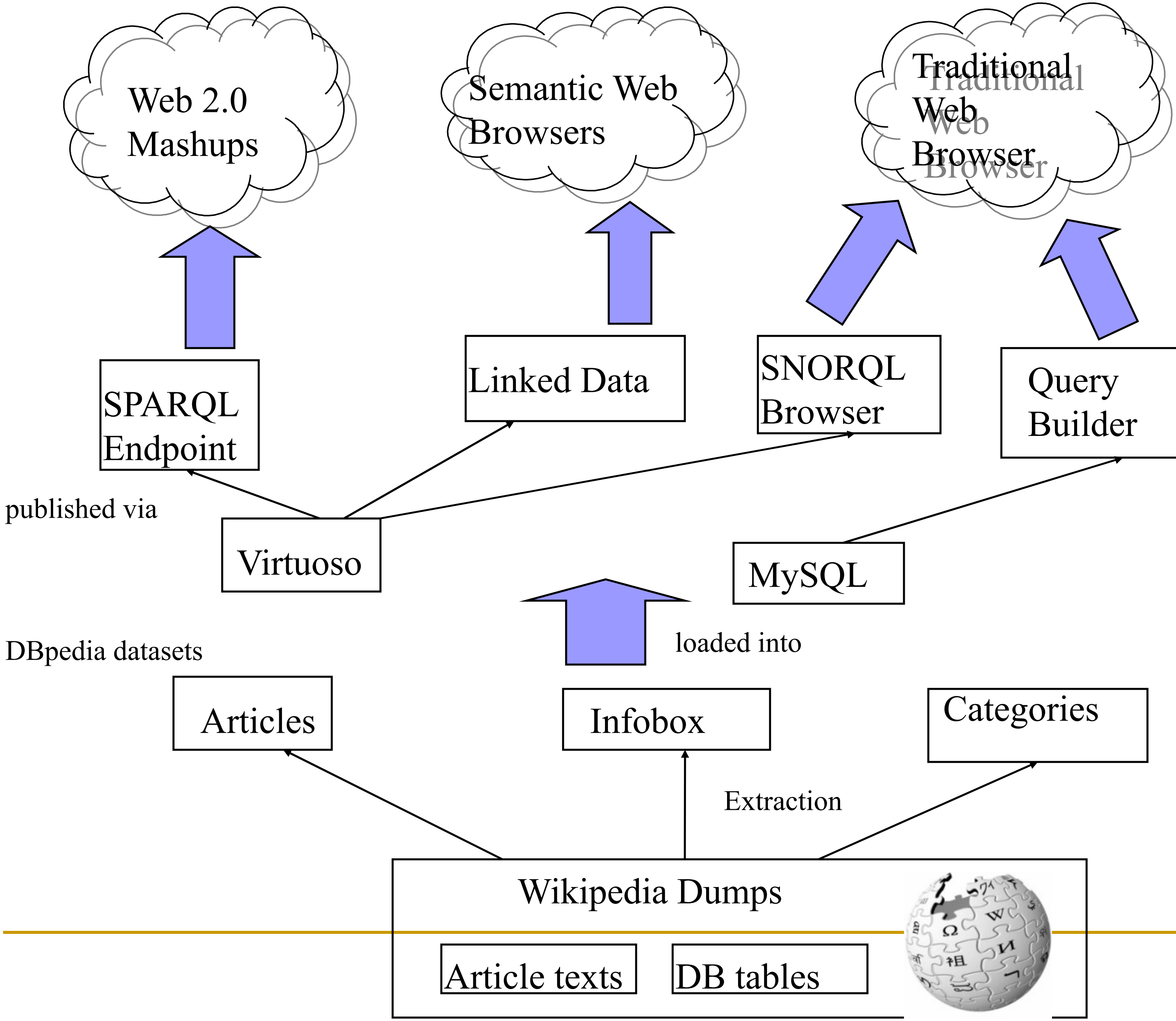
DBPedia – a further large scale knowledge source from Wikipedia

- This project analyses Wikipedia's infoboxes and transforms their content into RDF triples.
- Major problem: infobox attributes/values are not standarized
  - separate templates for Infobox_film, Infobox Film, and Infobox film.
- Templates are parsed recursively by RE
  - extracted relations are taken as they are (no manually defined heuristic for verification)
- Wikipedia categories are treated as classes and articles as individuals

DBPedia - sizes

- 115,000 classes, 650,000 individuals, sharing ~8000 types of semantic relations
- 103M rdf triples
  - 60% are internal linsk dereived from Wikipedia's link structure
  - 15% directly from infoboxes

More details



Web 2.0 Mashups

Semantic Web Browsers

Traditional Web Browser

Traditional Web Browser

SPARQL Endpoint

Linked Data

SNORQL Browser

Query Builder

published via

Virtuoso

MySQL

loaded into

DBpedia datasets

Articles

Infobox

Categories

Extraction

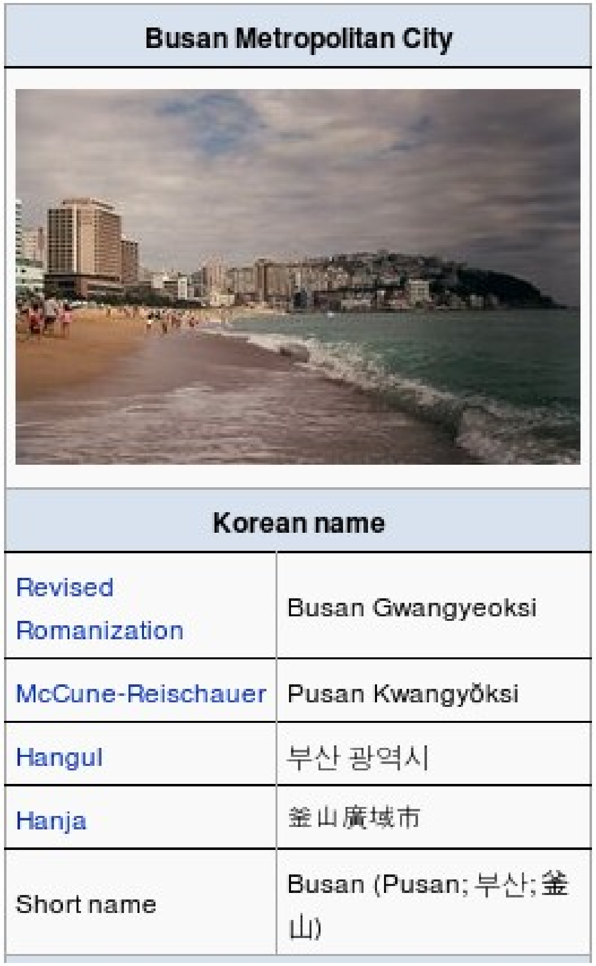
Wikipedia Dumps

Article texts

DB tables

Wikitext Syntax:

```
{{infobox City Korea|
  full_name=Busan Metropolitan City|
  image=[[Image:Haeundaebeachbusan.jpg|
    250px|Haeundae Beach, Busan]]|
  rr=Busan Gwangyeoksi|
  mr=Pusan Kwangyŏksi|
  hangul=부산 광역시|
  hanja=釜山廣域市|
  short_name=Busan (Pusan; 부산; 釜山)|
  population=3,635,389 ...|
  area=763.46 km²|
  government=[[Metropolitan cities of
    South Korea|Metropolitan City]]|
  divisions=15 wards (Gu),
    <br>1 county (Gun)|
  region=[[Yeongnam]]|
  dialect=[[Gyeongsang Dialect|
    Gyeongsang]]|
  map=[[Image:Busan map.png|Map of
    South Korea highlighting the city]]|
 }}
```

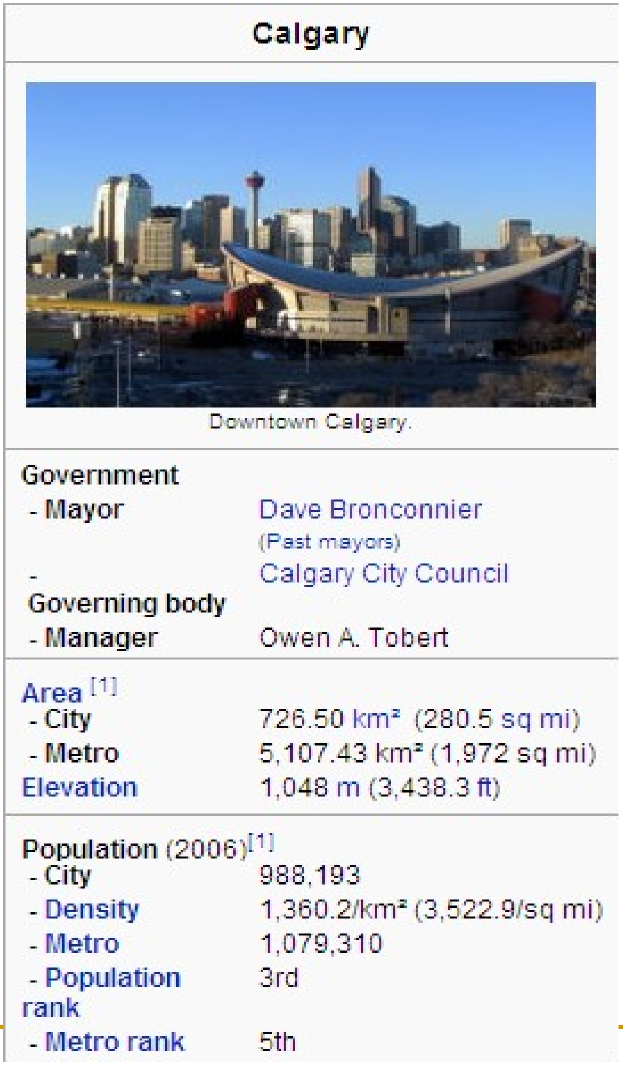


Busan Metropolitan City	
Korean name	
Revised Romanization	Busan Gwangyeoksi
McCune-Reischauer	Pusan Kwangyŏksi
Hangul	부산 광역시
Hanja	釜山廣域市
Short name	Busan (Pusan; 부산; 釜山)

Extracting Infobox Data (RDF Representation):

```
http://en.wikipedia.org/wiki/Calgary
```

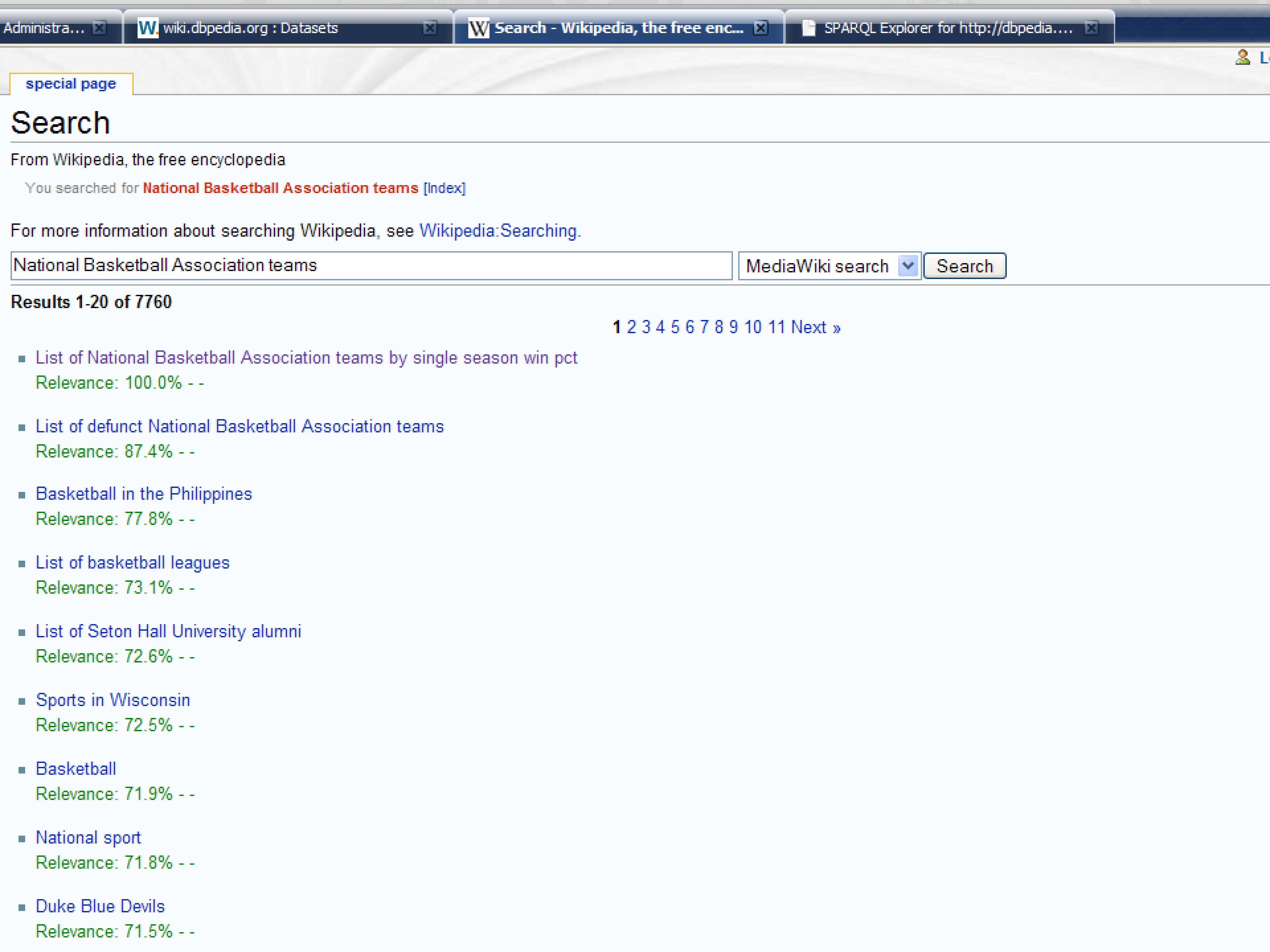
```
http://dbpedia.org/resource/Calgary
dbpedia:native_name Calgary”;
dbpedia:altitude “1048”;
dbpedia:population_city “988193”;
dbpedia:population_metro “1079310”;
mayor_name
      dbpedia:Dave_Bronconnier ;
governing_body
      dbpedia:Calgary_City_Council;
...
```



Calgary

Downtown Calgary.

Government	
- Mayor	Dave Bronconnier (Past mayors)
- Governing body	Calgary City Council
- Manager	Owen A. Tobert
Area [1]	
- City	726.50 km² (280.5 sq mi)
- Metro	5,107.43 km² (1,972 sq mi)
Elevation	1,048 m (3,438.3 ft)
Population (2006)[1]	
- City	988,193
- Density	1,360.2/km² (3,522.9/sq mi)
- Metro	1,079,310
- Population rank	3rd
- Metro rank	5th


Administra...
wiki.dbpedia.org : Datasets
Search - Wikipedia, the free enc...
SPARQL Explorer for http://dbpedia....
special page
MediaWiki search
Search

Search

From Wikipedia, the free encyclopedia

You searched for **National Basketball Association teams** [Index]

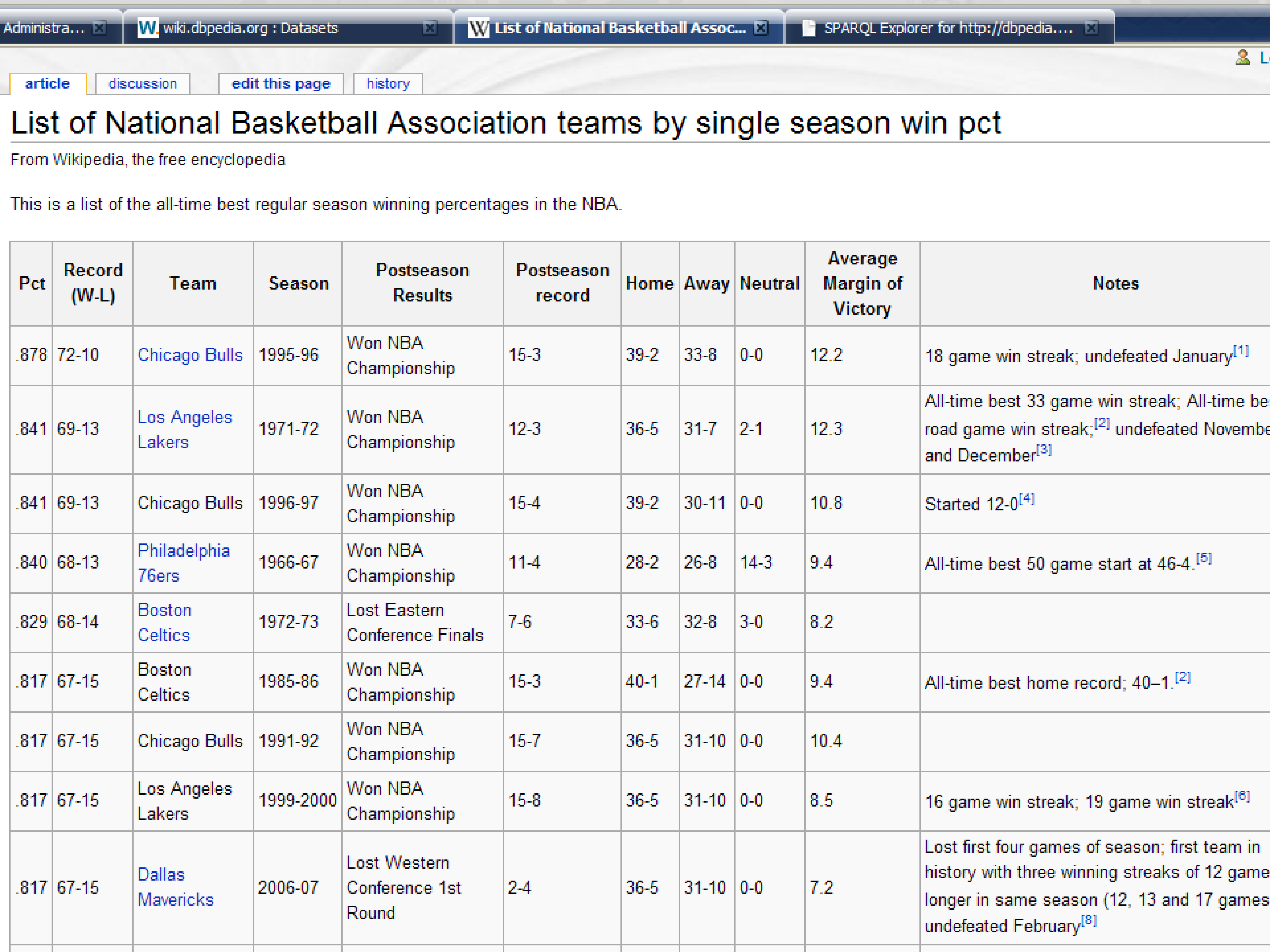
For more information about searching Wikipedia, see Wikipedia:Searching.

National Basketball Association teams

Results 1-20 of 7760

1 2 3 4 5 6 7 8 9 10 11 Next »

- List of National Basketball Association teams by single season win pct
  Relevance: 100.0% - -
- List of defunct National Basketball Association teams
  Relevance: 87.4% - -
- Basketball in the Philippines
  Relevance: 77.8% - -
- List of basketball leagues
  Relevance: 73.1% - -
- List of Seton Hall University alumni
  Relevance: 72.6% - -
- Sports in Wisconsin
  Relevance: 72.5% - -
- Basketball
  Relevance: 71.9% - -
- National sport
  Relevance: 71.8% - -
- Duke Blue Devils
  Relevance: 71.5% - -


Administra...
W. wiki.dbpedia.org : Datasets
List of National Basketball Assoc...
SPARQL Explorer for http://dbpedia....
article
discussion
edit this page
history

List of National Basketball Association teams by single season win pct

From Wikipedia, the free encyclopedia

This is a list of the all-time best regular season winning percentages in the NBA.

Pct	Record (W-L)	Team	Season	Postseason Results	Postseason record	Home	Away	Neutral	Average Margin of Victory	Notes
.878	72-10	Chicago Bulls	1995-96	Won NBA Championship	15-3	39-2	33-8	0-0	12.2	18 game win streak; undefeated January[1]
.841	69-13	Los Angeles Lakers	1971-72	Won NBA Championship	12-3	36-5	31-7	2-1	12.3	All-time best 33 game win streak; All-time be road game win streak;[2] undefeated Novembe and December[3]
.841	69-13	Chicago Bulls	1996-97	Won NBA Championship	15-4	39-2	30-11	0-0	10.8	Started 12-0[4]
.840	68-13	Philadelphia 76ers	1966-67	Won NBA Championship	11-4	28-2	26-8	14-3	9.4	All-time best 50 game start at 46-4.[5]
.829	68-14	Boston Celtics	1972-73	Lost Eastern Conference Finals	7-6	33-6	32-8	3-0	8.2	
.817	67-15	Boston Celtics	1985-86	Won NBA Championship	15-3	40-1	27-14	0-0	9.4	All-time best home record; 40–1.[2]
.817	67-15	Chicago Bulls	1991-92	Won NBA Championship	15-7	36-5	31-10	0-0	10.4	
.817	67-15	Los Angeles Lakers	1999-2000	Won NBA Championship	15-8	36-5	31-10	0-0	8.5	16 game win streak; 19 game win streak[6]
.817	67-15	Dallas Mavericks	2006-07	Lost Western Conference 1st Round	2-4	36-5	31-10	0-0	7.2	Lost first four games of season; first team in history with three winning streaks of 12 game longer in same season (12, 13 and 17 games undefeated February[8]

DBpedia Basics :

The structured information can be extracted from Wikipedia and can serve as a basis for enabling sophisticated queries against Wikipedia content.

The DBpedia.org project uses the Resource Description Framework (RDF) as a flexible data model for representing extracted information and for publishing it on the Web. It uses the SPARQL query language to query this data. At Developers Guide to Semantic Web Toolkits you find a development toolkit in your preferred programming language to process DBpedia data.

Accessing the DBpedia Dataset over the Web

1. SPARQL Endpoint
2. Linked Data Interface
3. DB Dumps for Download

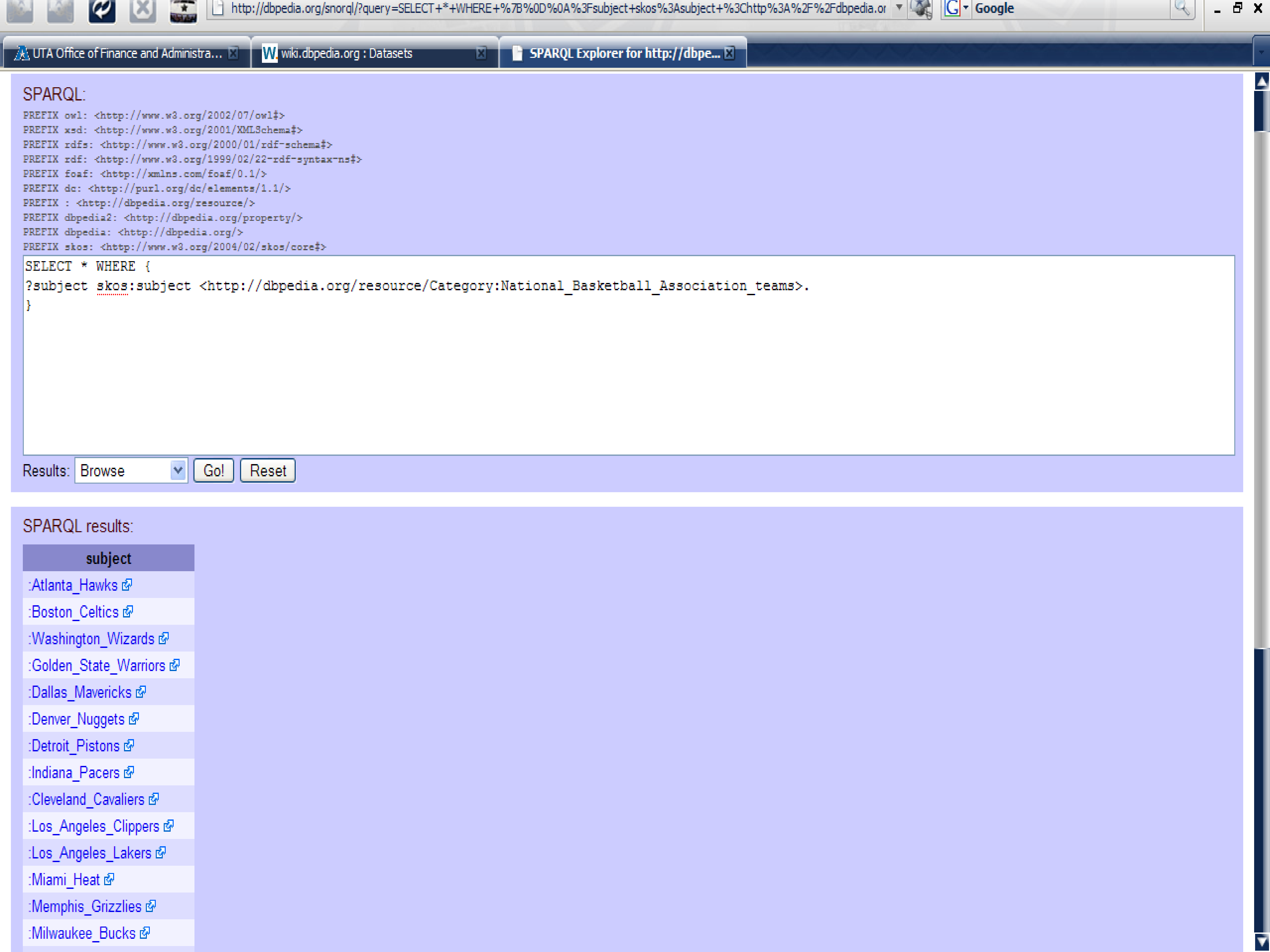
SPARQL :

- SPARQL is a query language for RDF.
- RDF is a directed, labeled graph data format for representing information in the Web.
- This specification defines the syntax and semantics of the SPARQL query language for RDF.
- SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware.

The DBpedia SPARQL Endpoint

- http://dbpedia.org/sparql
- hosted on a OpenLink Virtuoso server
- can answer SPARQL queries like
  - Give me all Sitcoms that are set in NYC?
  - All tennis players from Moscow?
  - All films by Quentin Tarentino?
  - All German musicians that were born in Berlin in the 19th century?
- Provides two extensions to SPARQL
  - free-text search within titles and abstracts
  - COUNT()

*SPARQL wasn't working so all the following examples are from SNORQL


http://dbpedia.org/snorql/?query=SELECT+*+WHERE+%7B%0D%0A%3Fsubject+skos%3Asubject+%3Chttp%3A%2F%2Fdbpedia.or
Google
UTA Office of Finance and Administra...
wiki.dbpedia.org : Datasets
SPARQL Explorer for http://dbpe...

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT * WHERE {
?subject skos:subject <http://dbpedia.org/resource/Category:National_Basketball_Association_teams>.
}
```

Results: Browse

SPARQL results:

subject
:Atlanta_Hawks
:Boston_Celtics
:Washington_Wizards
:Golden_State_Warriors
:Dallas_Mavericks
:Denver_Nuggets
:Detroit_Pistons
:Indiana_Pacers
:Cleveland_Cavaliers
:Los_Angeles_Clippers
:Los_Angeles_Lakers
:Miami_Heat
:Memphis_Grizzlies
:Milwaukee_Bucks

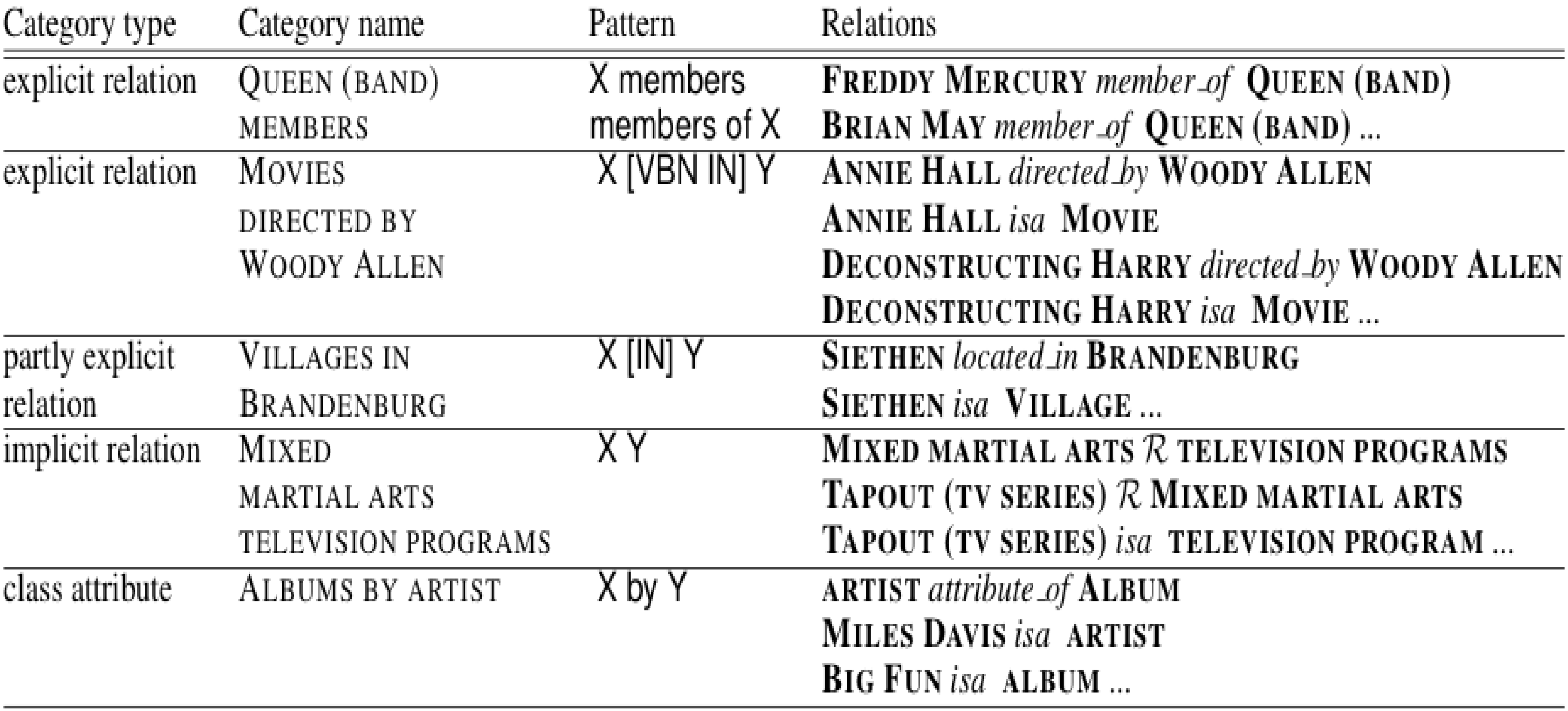
EMLR – Mining for knowledge in Wikipedia categories

- Observations, cf. Nastase & Strube, 2008
  - Wikipedia categories have complex names
    - Reflecting human classification & organization instances
    - Implicitliy, encode knowledge about class attributes, taxonomic and other semantic relations
- Goal:
  - Extract this implicit knowledge
  - Use it for creating structured knowledge base

Examples of Wikipedia categories

- Books by Genre
  - Children's books, reference work, textbooks, Novels
- Newspapers published by NewsQuest
  - Evening Times, The Oxford Times
- Goal:
  - Develop methods that automatically decode thins strings and determine the relations, classes and attributes they encode.

Categorie Names and the encoding relations



Category type	Category name	Pattern	Relations
explicit relation	Queen (band) members	X members members of X	Freddy Mercury *member_of* Queen (band) Brian May *member_of* Queen (band) ...
explicit relation	Movies directed by Woody Allen	X [VBN IN] Y	Annie Hall *directed_by* Woody Allen Annie Hall *isa* Movie Deconstructing Harry *directed_by* Woody Allen Deconstructing Harry *isa* Movie ...
partly explicit relation	Villages in Brandenburg	X [IN] Y	Siethen *located_in* Brandenburg Siethen *isa* Village ...
implicit relation	Mixed martial arts television programs	X Y	Mixed martial arts $\mathcal{R}$ television programs Tapout (TV series) $\mathcal{R}$ Mixed martial arts Tapout (TV series) *isa* television program ...
class attribute	Albums by artist	X by Y	artist *attribute_of* Album Miles Davis *isa* artist Big Fun *isa* album ...

Table 1: Examples of information encoded in category names and the knowledge we extract

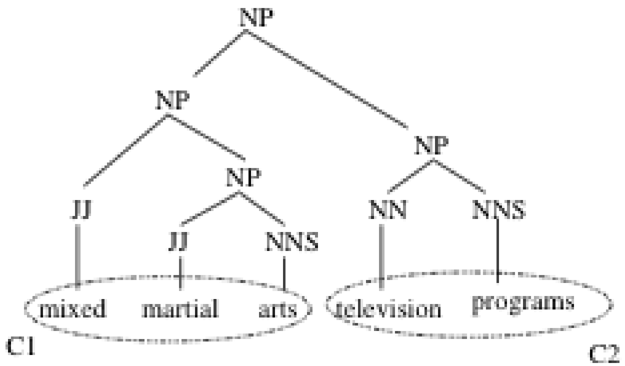
Import step: syntactic analysis of categorie

- Identify phrase structure of noun compound
- Identify dominant constitiuent
  - Chairmen For The County Councils Of Norway
    - 3 constituents: chrairmen, county council, Norway
    - Dominating constituent: chairmen

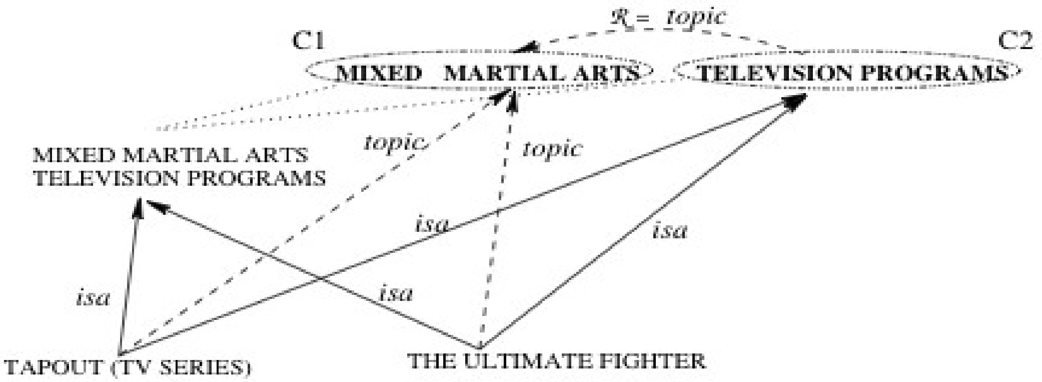
Extracting explicit relations

- Explicit relaton
  - Queen (band) members → memberOf(P, X)
    - memberOf(Brian Mary, Queen)
- Categorie title
  - Movie directed by Woody Allen
    - X [VBN IN] Y → isA(P, X)
- Partially explicit relation
  - X [IN] Y
    - If X=Person & Y=Organization → isA(P, X) & memberOf(P, Y)
    - If Y=LOC → isA(P,X) & spatial(P, Y)

Extract implicit relations

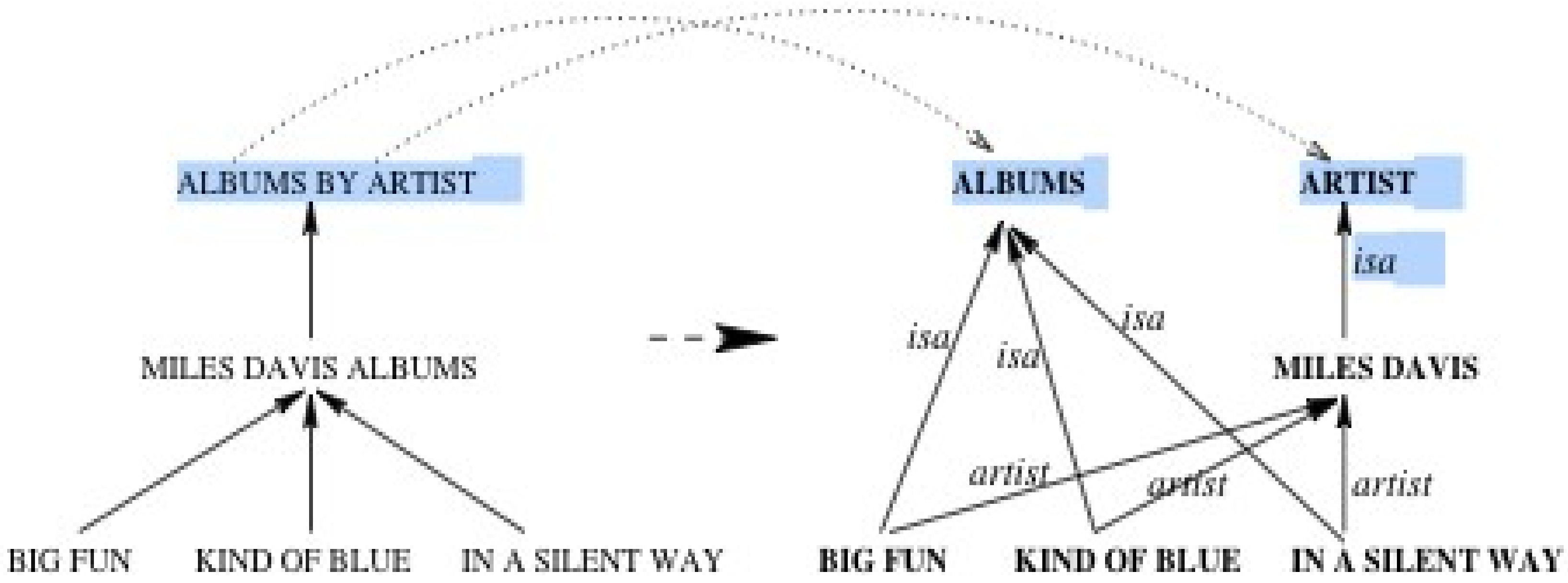


- If categories are complex nouns, do NC analysis
- Propagate extracted relations R to corresponding pages



Extraction of class attributes and attribute values

- „By"-cases, e.g., albums by Miles Davis



Results

- Sizes:
  - 3.4M isA, 3.2M spatial
  - 43,000 memberOf, 44,000 other relation (causedBy, writtenBy)
- 4 samples of 250 relations by humans
  - 84%-98% precision

Intermediate Summary

- Yago, DBPedia, EMLR extracted knowledge bases
- Large scale
- Difficult to compare, because extracted relations differ
  - WrittenInYear → Yaho
  - WrittenBy → EMLR
  - Written, writtenBy, writer, writers, writerName, coWriter → DBPedia
- However, do play important role in large-scale Semantic Web → linked data → see later