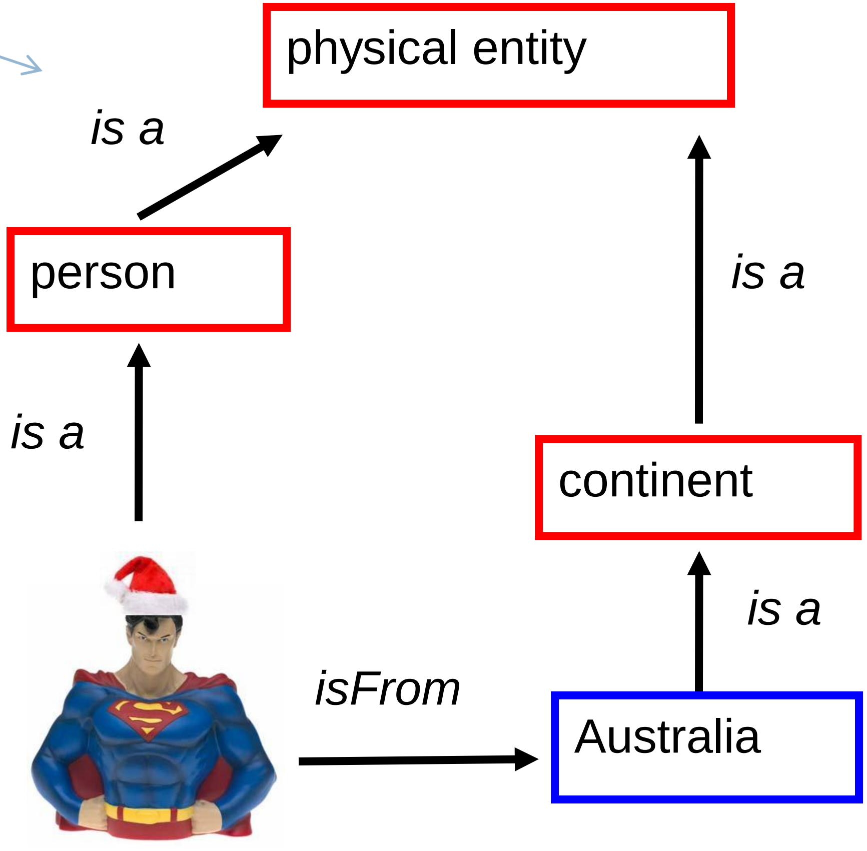
YAGO: A LARGE ONTOLOGY FROM WIKIPEDIA AND WORDNET

Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weiku
Web Sem. 6(3): 203-217 (2008)

Presented by,
Quazi Mainul Hasan
1000629641
CS Dept. UT Arlington.

Background

- Ontology



physical entity

is a

person

is a

is a

continent

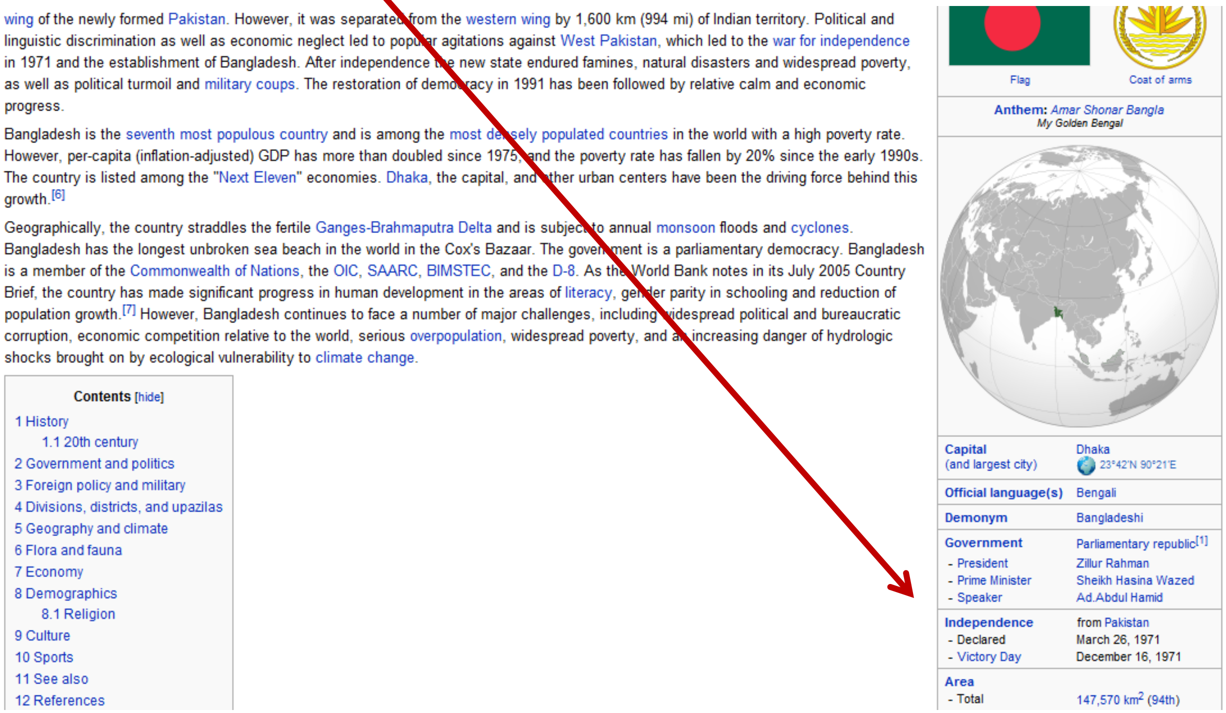
is a

isFrom

Australia

Background

- Ontology
- Infobox in Wikipedia



wing of the newly formed Pakistan. However, it was separated from the western wing by 1,600 km (994 mi) of Indian territory. Political and linguistic discrimination as well as economic neglect led to popular agitations against West Pakistan, which led to the war for independence in 1971 and the establishment of Bangladesh. After independence the new state endured famines, natural disasters and widespread poverty, as well as political turmoil and military coups. The restoration of democracy in 1991 has been followed by relative calm and economic progress.

Bangladesh is the seventh most populous country and is among the most densely populated countries in the world with a high poverty rate. However, per-capita (inflation-adjusted) GDP has more than doubled since 1975, and the poverty rate has fallen by 20% since the early 1990s. The country is listed among the "Next Eleven" economies. Dhaka, the capital, and other urban centers have been the driving force behind this growth.[6]

Geographically, the country straddles the fertile Ganges-Brahmaputra Delta and is subject to annual monsoon floods and cyclones. Bangladesh has the longest unbroken sea beach in the world in the Cox's Bazaar. The government is a parliamentary democracy. Bangladesh is a member of the Commonwealth of Nations, the OIC, SAARC, BIMSTEC, and the D-8. As the World Bank notes in its July 2005 Country Brief, the country has made significant progress in human development in the areas of literacy, gender parity in schooling and reduction of population growth.[7] However, Bangladesh continues to face a number of major challenges, including widespread political and bureaucratic corruption, economic competition relative to the world, serious overpopulation, widespread poverty, and an increasing danger of hydrologic shocks brought on by ecological vulnerability to climate change.

Contents [hide]

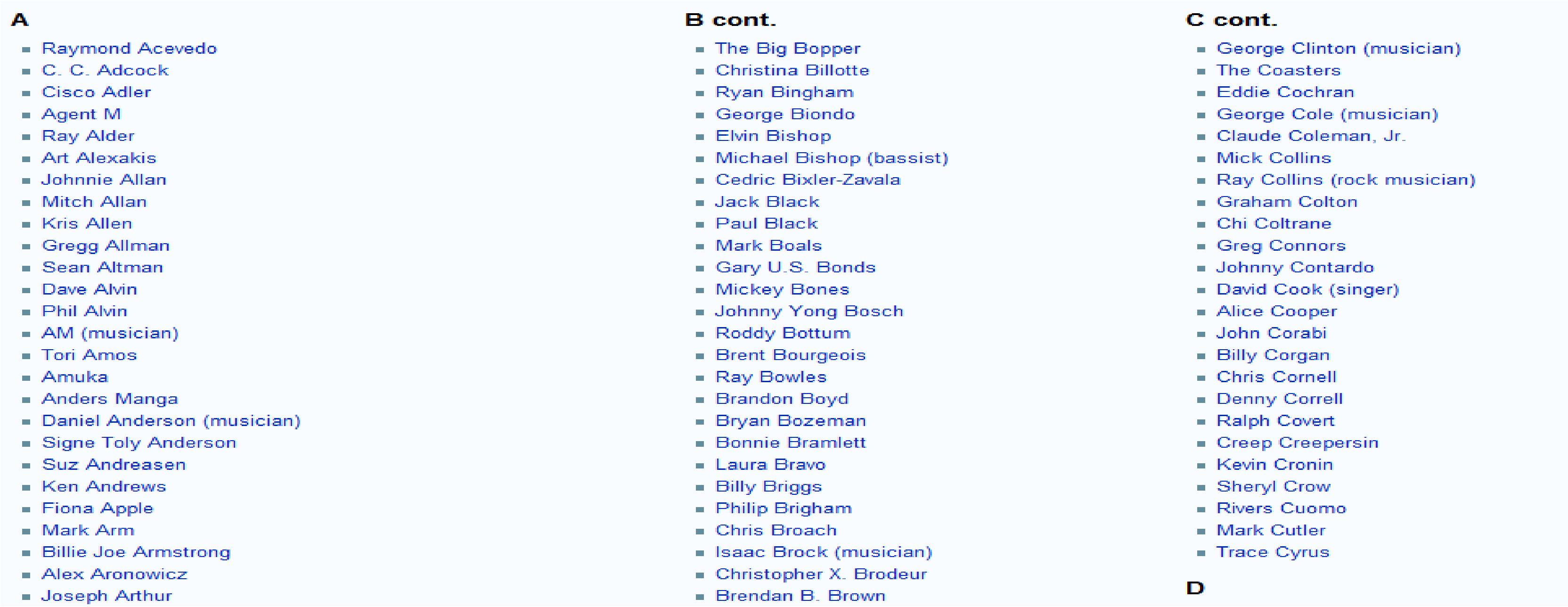
Flag Coat of arms

Anthem: *Amar Shonar Bangla*
My Golden Bengal

Capital (and largest city)	Dhaka 23°42′N 90°21′E
Official language(s)	Bengali
Demonym	Bangladeshi
Government	Parliamentary republic[1]
- President	Zillur Rahman
- Prime Minister	Sheikh Hasina Wazed
- Speaker	Ad.Abdul Hamid
Independence	from Pakistan
- Declared	March 26, 1971
- Victory Day	December 16, 1971
Area	
- Total	147,570 km^2 (94th)

Background

- Ontology
- Infobox in Wikipedia
- Wiki category pages



A

- Raymond Acevedo
- C. C. Adcock
- Cisco Adler
- Agent M
- Ray Alder
- Art Alexakis
- Johnnie Allan
- Mitch Allan
- Kris Allen
- Gregg Allman
- Sean Altman
- Dave Alvin
- Phil Alvin
- AM (musician)
- Tori Amos
- Amuka
- Anders Manga
- Daniel Anderson (musician)
- Signe Toly Anderson
- Suz Andreasen
- Ken Andrews
- Fiona Apple
- Mark Arm
- Billie Joe Armstrong
- Alex Aronowicz
- Joseph Arthur

B cont.

- The Big Bopper
- Christina Billotte
- Ryan Bingham
- George Biondo
- Elvin Bishop
- Michael Bishop (bassist)
- Cedric Bixler-Zavala
- Jack Black
- Paul Black
- Mark Boals
- Gary U.S. Bonds
- Mickey Bones
- Johnny Yong Bosch
- Roddy Bottum
- Brent Bourgeois
- Ray Bowles
- Brandon Boyd
- Bryan Bozeman
- Bonnie Bramlett
- Laura Bravo
- Billy Briggs
- Philip Brigham
- Chris Broach
- Isaac Brock (musician)
- Christopher X. Brodeur
- Brendan B. Brown

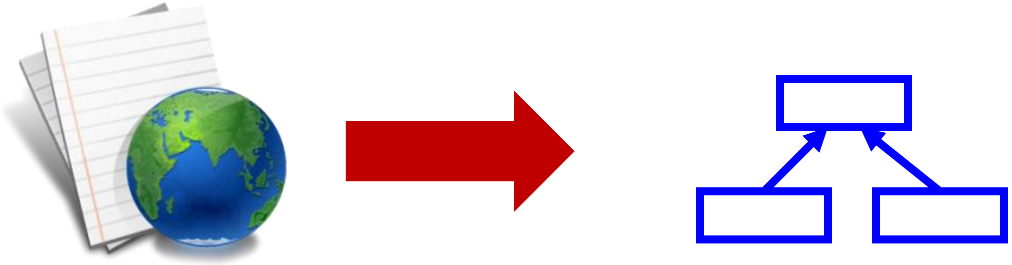
C cont.

- George Clinton (musician)
- The Coasters
- Eddie Cochran
- George Cole (musician)
- Claude Coleman, Jr.
- Mick Collins
- Ray Collins (rock musician)
- Graham Colton
- Chi Coltrane
- Greg Connors
- Johnny Contardo
- David Cook (singer)
- Alice Cooper
- John Corabi
- Billy Corgan
- Chris Cornell
- Denny Correll
- Ralph Covert
- Creep Creepersin
- Kevin Cronin
- Sheryl Crow
- Rivers Cuomo
- Mark Cutler
- Trace Cyrus

D

Vision

- Gathering the knowledge of this world in a structured ontology.



1. Semantic Search

2. Question answering

Approach

- Extract candidate entities and facts from Wikipedia in connection with WordNet
- Use extensive quality control techniques

Yago Model Concepts

- All objects are Entities
- Words are also entities
- Similar Entities are grouped into classes
- Each entity is an instance of at least one class
- Classes are entities too
- Relationships are also entities

Elvis won a Grammy Award -> Elvis Presley **HASWONPRIZE** Grammy Award

“Elvis” MEANS Elvis Presley

“Elvis” MEANS Elvis Costello

Elvis Presley TYPE Singer

singer SUBCLASSOF Person

Subclassof TYPE atr

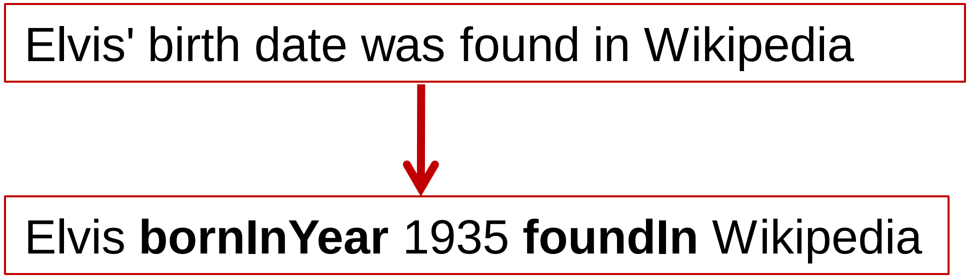
Yago Model Concepts contd.

- <entity, relation, entity> = fact
- Fact are identified with a fact identifier

(Elvis Presley, BORNINYEAR, 1935)= indentifier #1

- Each fact is stored with it’s location

#1 FOUNDIN Wikipedia

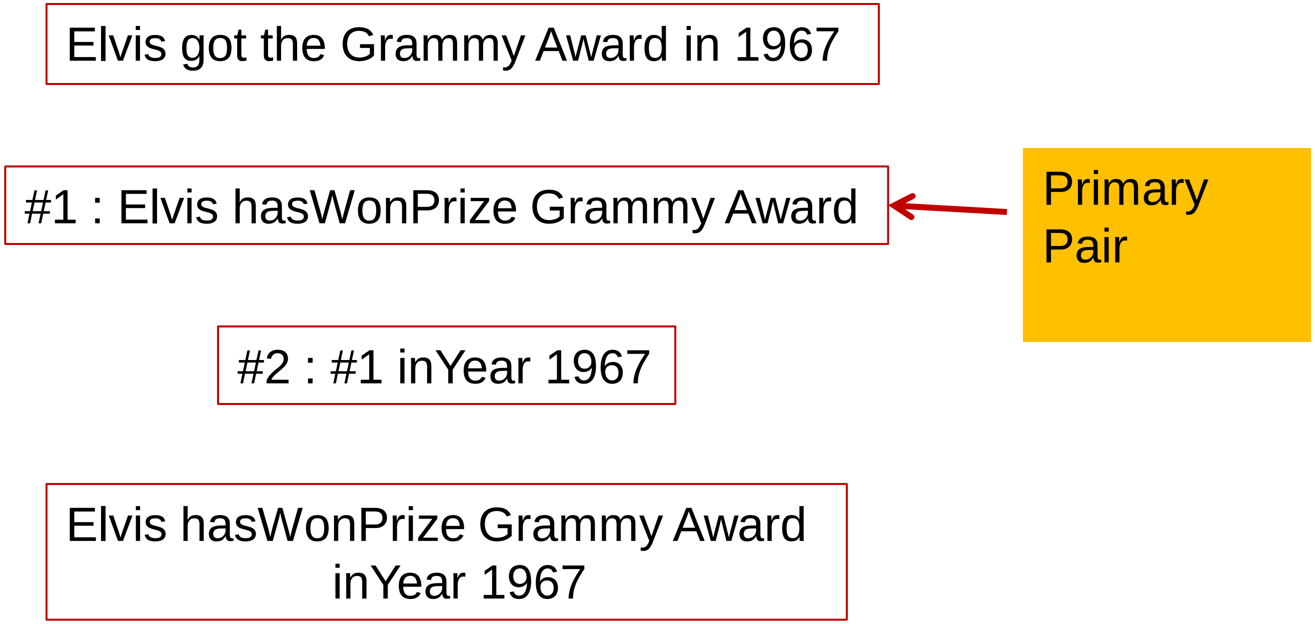


Elvis' birth date was found in Wikipedia

Elvis **bornInYear** 1935 **foundIn** Wikipedia

n-ary relations

- Facts with more than two arguments



Elvis got the Grammy Award in 1967

#1 : Elvis hasWonPrize Grammy Award

Primary Pair

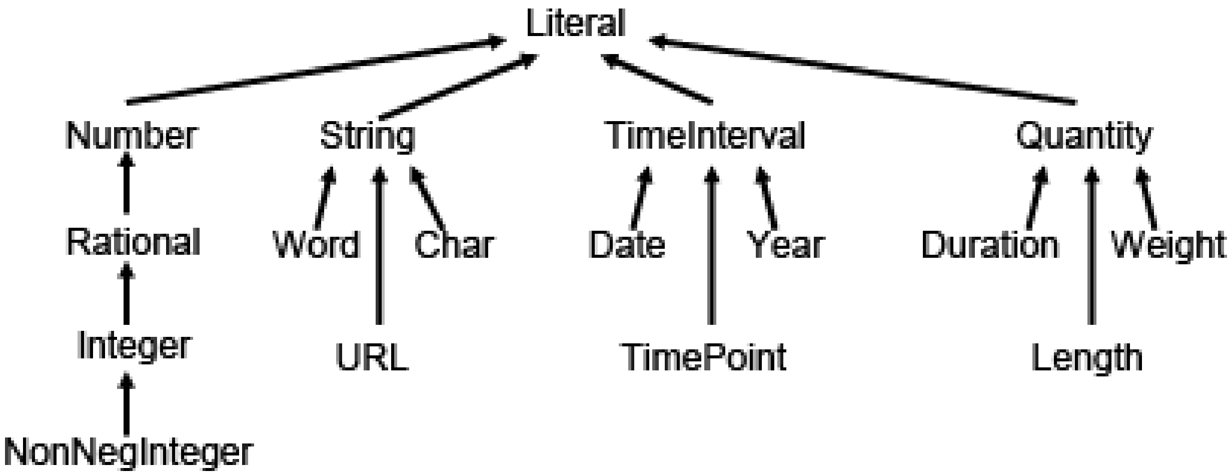
#2 : #1 inYear 1967

Elvis hasWonPrize Grammy Award inYear 1967

Other Concepts

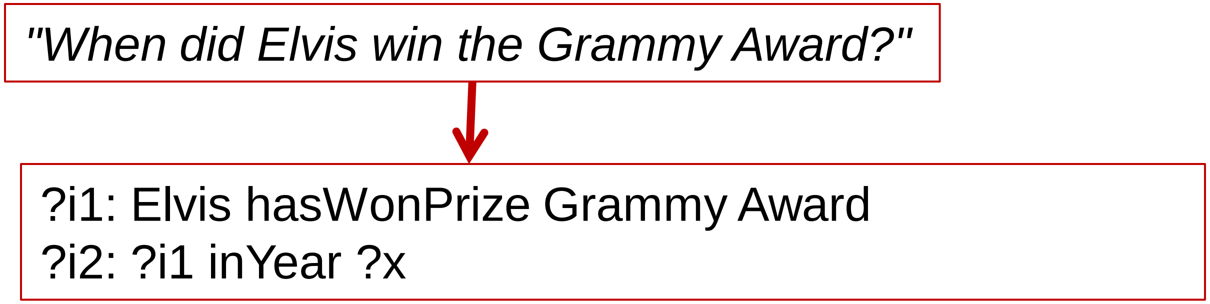
□ Data Types

1. Treats literals as proper entities
2. Literals are instances of literal classes



Query Language

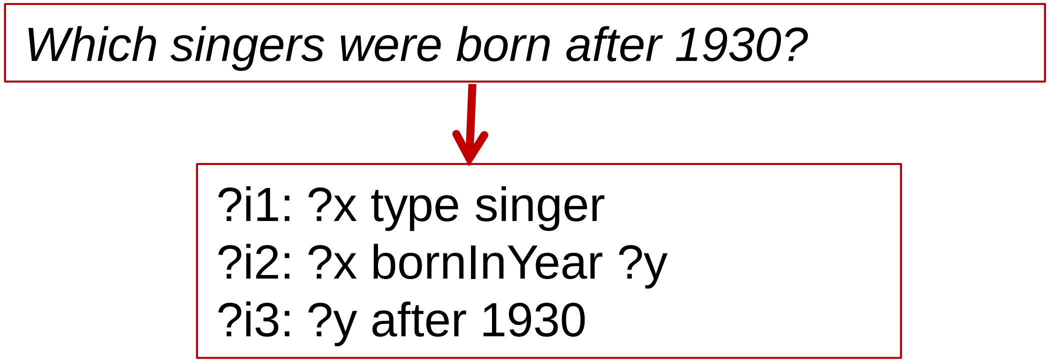
- Demonstrates the use of YAGO



"When did Elvis win the Grammy Award?"

?i1: Elvis hasWonPrize Grammy Award
?i2: ?i1 inYear ?x

- Filter Relations: BEORE or AFTER

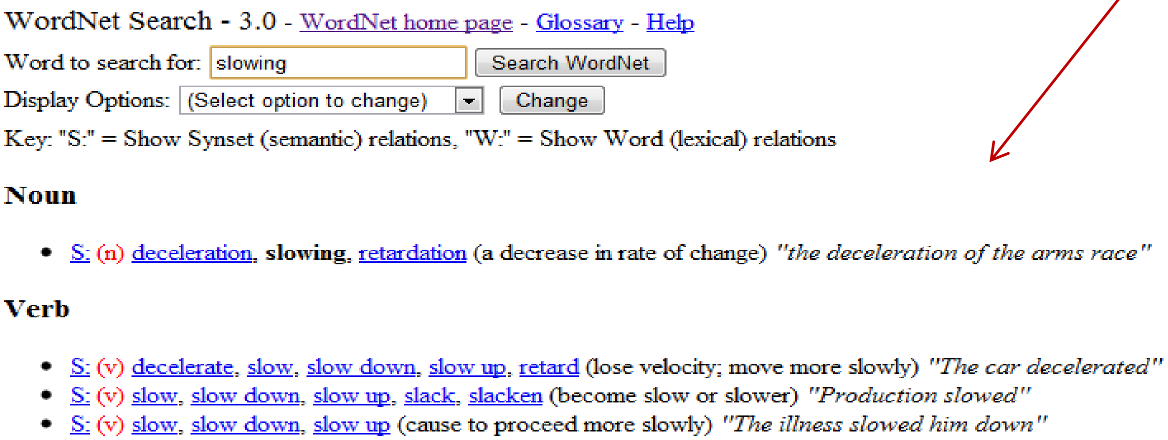


Which singers were born after 1930?

?i1: ?x type singer
?i2: ?x bornInYear ?y
?i3: ?y after 1930

Assumption based on WordNet

- Distinguishes between words and actual senses of the words.
- Synset – set of words share one sense
- Only Nouns are considered here.
- Focused on hyponyms



WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: slowing

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) deceleration, **slowing**, retardation (a decrease in rate of change) *"the deceleration of the arms race"*

Verb

- S: (v) decelerate, slow, slow down, slow up, retard (lose velocity; move more slowly) *"The car decelerated"*
- S: (v) slow, slow down, slow up, slack, slacken (become slow or slower) *"Production slowed"*
- S: (v) slow, slow down, slow up (cause to proceed more slowly) *"The illness slowed him down"*

Assumption based on Wikipedia

- Each wiki article is an entity
- Each entity is assigned categories
- Infobox contains information about an entity in a standardized table
- People contains birthdates, profession and nationality
- XML Dump of wiki is used.

Infobox Heuristics

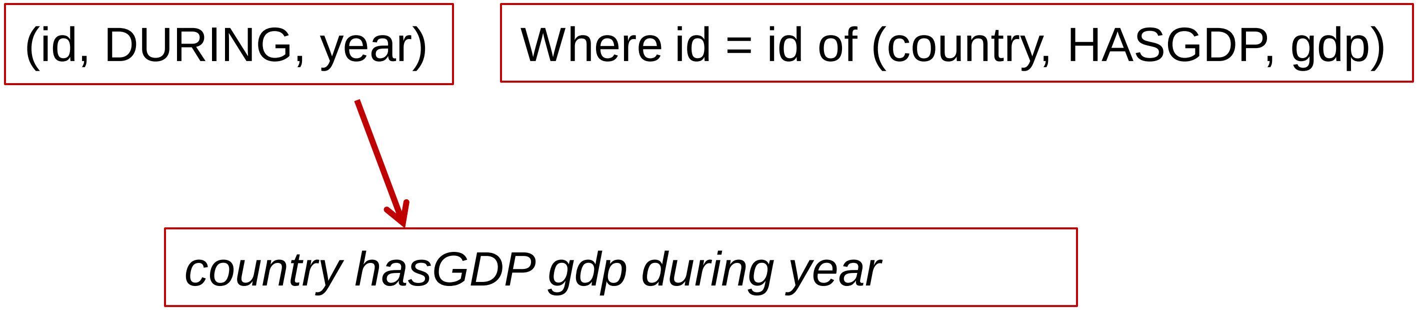
- □ Mapping from an attribute to a target relation

  BORN -> BIRTHDATE

- □ Whether the attributes is inverse attribute

  Official name, MEANS, entity

- □ Whether it allows multiple values
- □ Whether it is about another fact



(id, DURING, year)

Where id = id of (country, HASGDP, gdp)

country hasGDP gdp during year

Type Heuristics

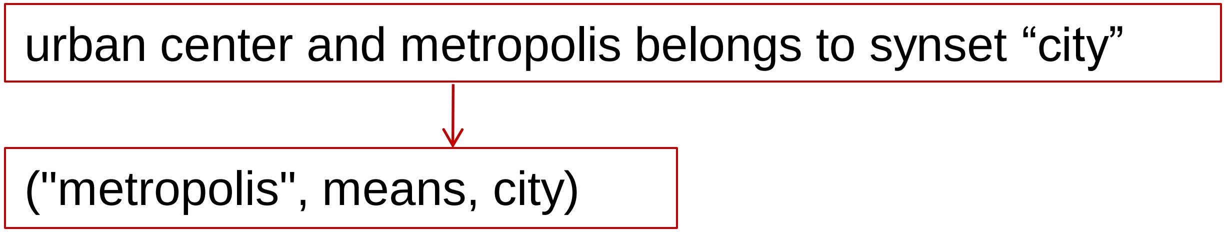
- Different types of categories
- Conceptual category

  Albert Einstein is in category ***Naturalized citizens of the United States***

- Shallow linguistic parsing
  1. Pre-modifier, a head and post-modifier
  2. If a head is plural, it is conceptual category
- Pling-Stemmer to identify and stem plural word

Type Heuristics(contd)

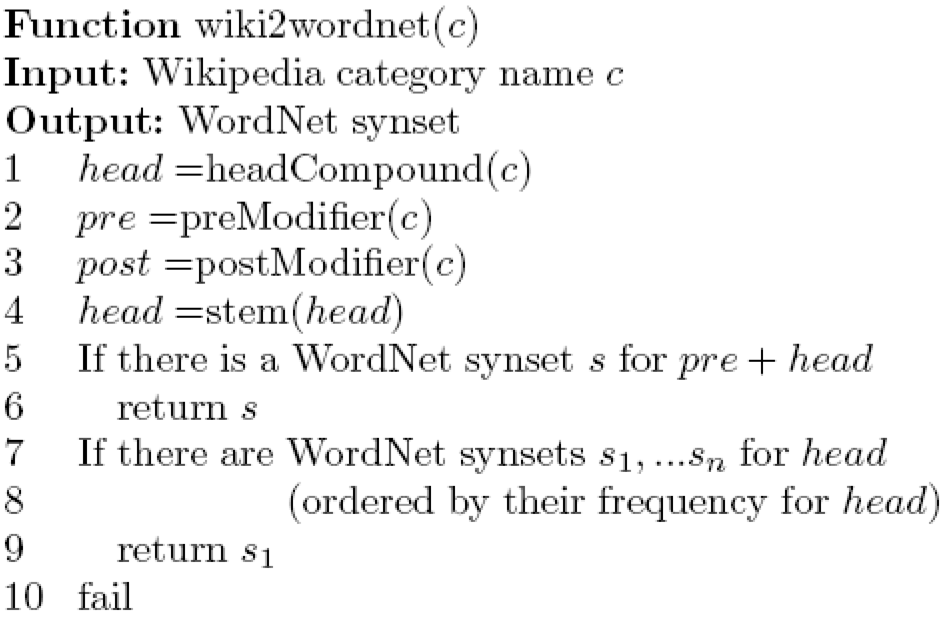
- Leafs categories are considered from Wikipedia
- WordNet is used to establish the hierarchy of classes
- Word Heuristics
  - Each synset becomes a class of YAGO



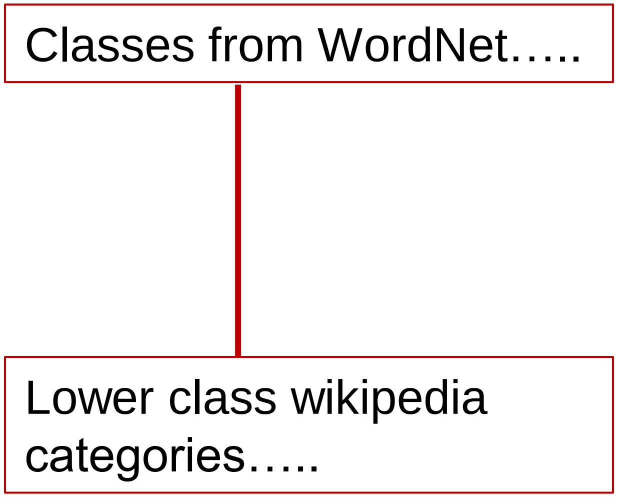
urban center and metropolis belongs to synset “city”

("metropolis", means, city)

Connecting Wikipedia and WordNet



```
Function wiki2wordnet($c$)
Input: Wikipedia category name $c$
Output: WordNet synset
$head$ =headCompound($c$)
$pre$ =preModifier($c$)
$post$ =postModifier($c$)
$head$ =stem($head$)
If there is a WordNet synset $s$ for $pre + head$
    return $s$
If there are WordNet synsets $s_1, ...s_n$ for $head$
                (ordered by their frequency for $head$)
    return $s_1$
fail
```

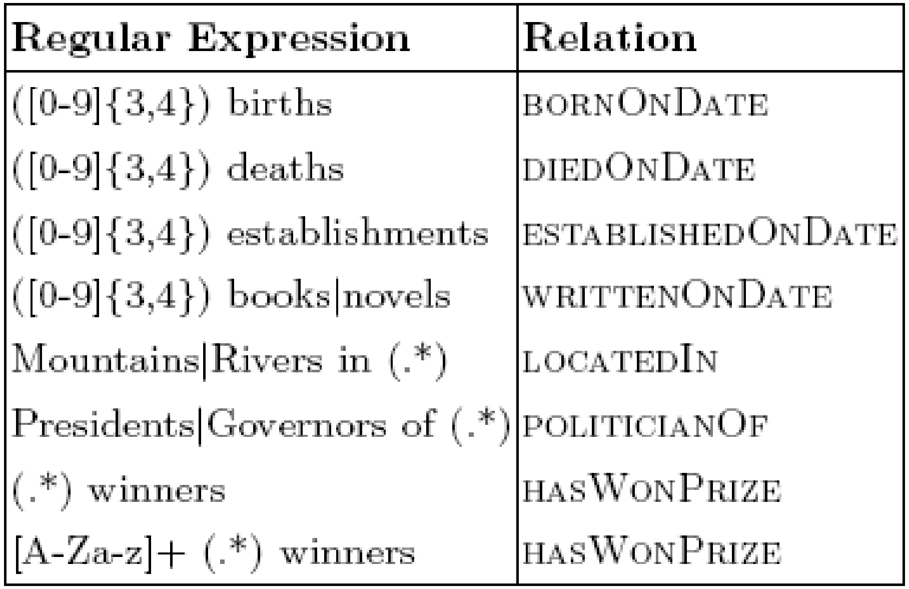


Classes from WordNet…..

Lower class wikipedia categories…..

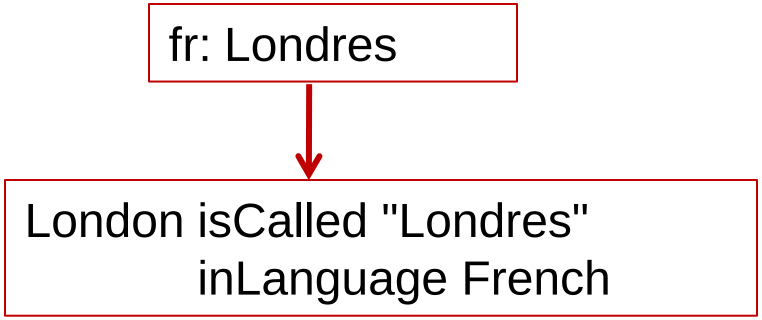
Category Heuristics

- Relation categories
  - Regular expression is used.



Regular Expression	Relation
([0-9]{3,4}) births	bornOnDate
([0-9]{3,4}) deaths	diedOnDate
([0-9]{3,4}) establishments	establishedOnDate
([0-9]{3,4}) books|novels	writtenOnDate
Mountains|Rivers in (.*)	locatedIn
Presidents|Governors of (.*)	politicianOf
(.*) winners	hasWonPrize
[A-Za-z]+ (.*) winners	hasWonPrize

- Language categories

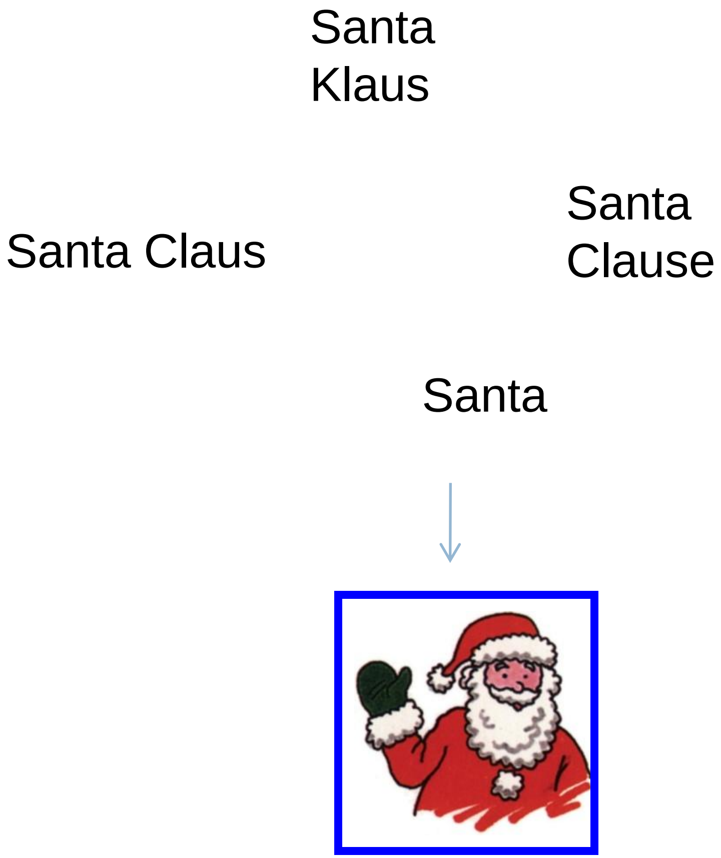


fr: Londres

London isCalled "Londres" inLanguage French

Quality Control

1. Canonicalization
   1. **Redirect Resolution**



Santa Klaus

Santa Claus

Santa Clause

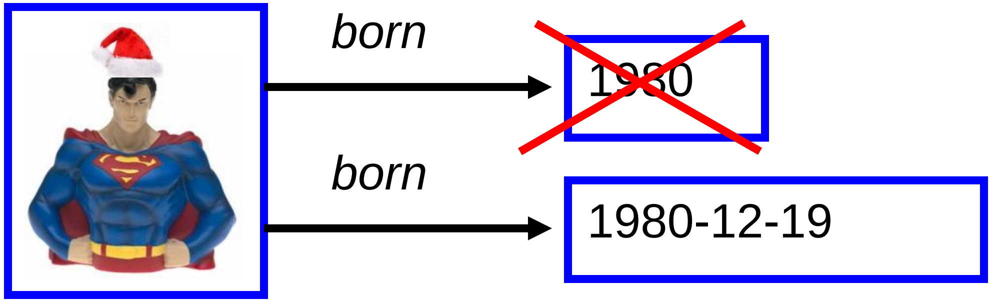
Santa

Quality Control

1. Canonicalization

1.1. Redirect Resolution

1. 2. **Duplicate Fatcs removal**



born

~~1980~~

born

1980-12-19

Quality Control

1. Canonicalization

1.1. Redirect Resolution

1. 2. Duplicate Fatcs removal

2. Type Checking

2.1 Reductive type Checking

2.2 Inductive Type Checking

range(bornOnDate, timepoint)

~~bornOnDate(Claus_Kent, Sydney)~~

Quality Control

1. Canonicalization

1.1. Redirect Resolution

1. 2. Duplicate Fatcs removal

Every fact and every entity
occurs exactly once

2. Type Checking

2.1 Reductive type Checking

2.2 Inductive Type Checking

Every fact fulfills
its type constraints

entity with Birth date -> person
instead of deleting it.

Storage

- DESCRIBE relation between individual and it's URL

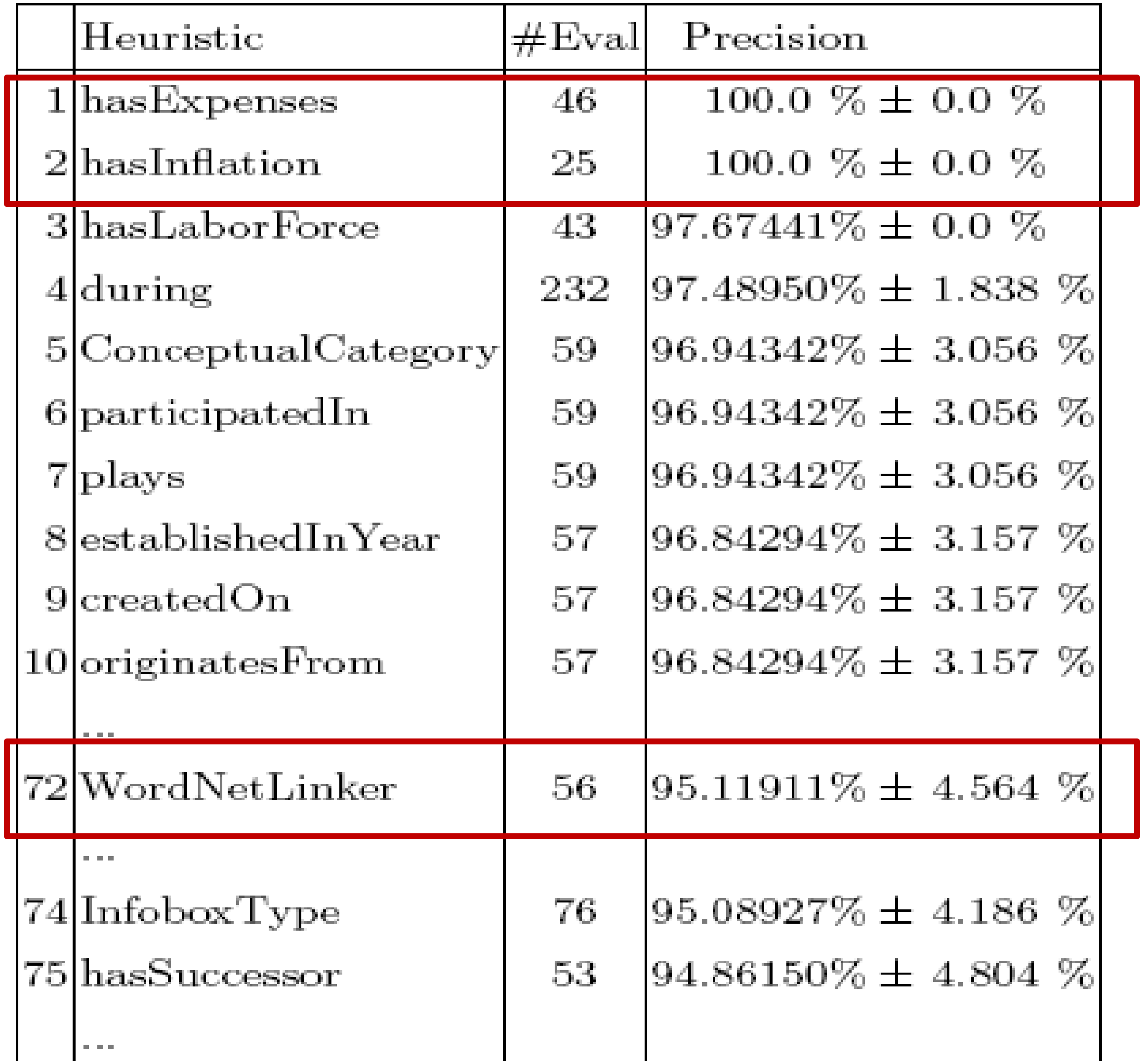
  Albert Einstein DESCRIBES http://en.wikipedia.org/wiki/Albert_Einstein

- Witness – USING, FOUNDIN, DURING
- FileFormat

  FACTS(factid, arg1, realtion, arg2, accuracy)

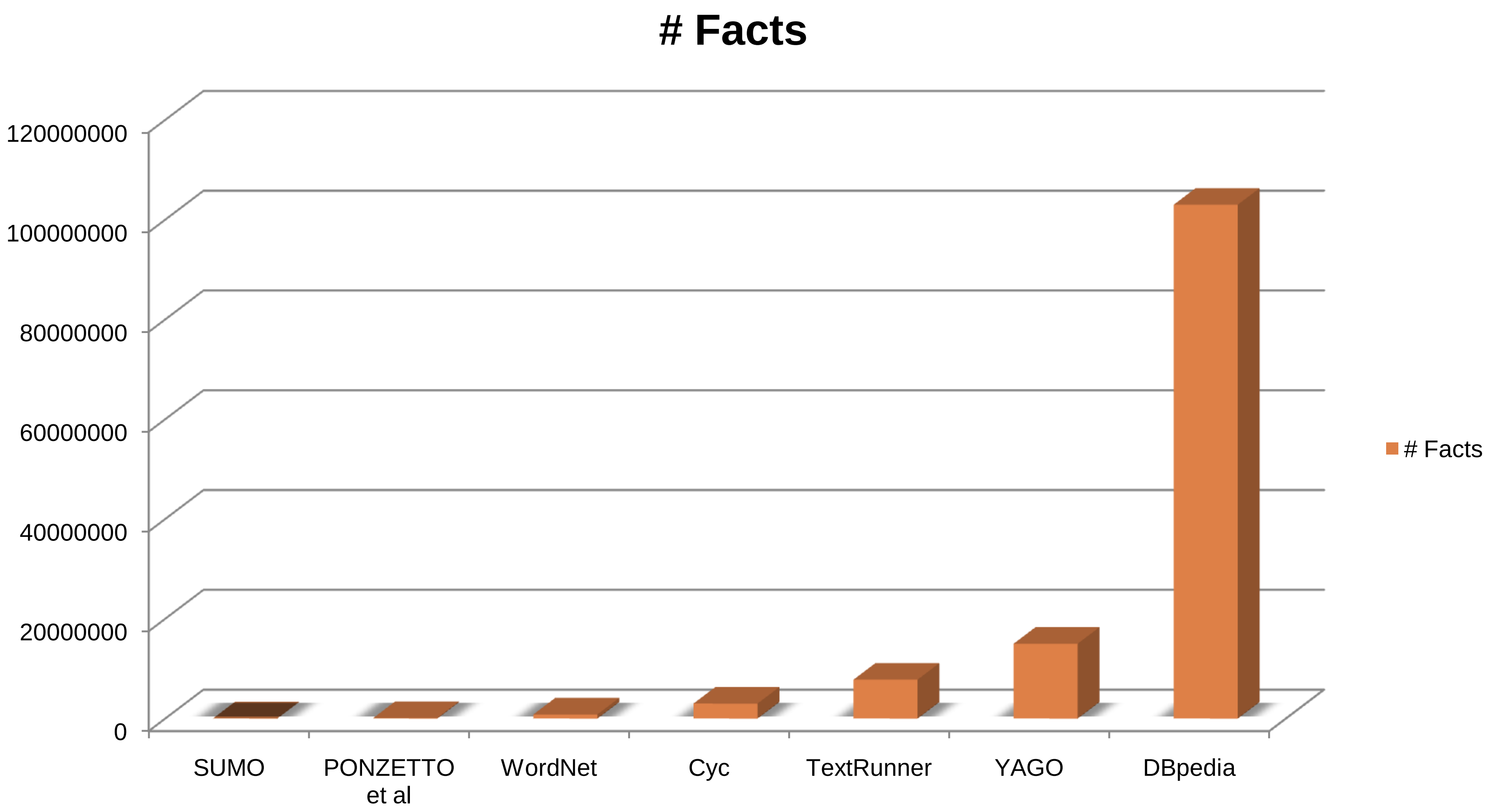
Evaluation

- Manual evaluation for ontology precision
- 13 judges evaluates 5200 facts
- YAGO includes 92 relations, 224391 classes and 1531588 individuals

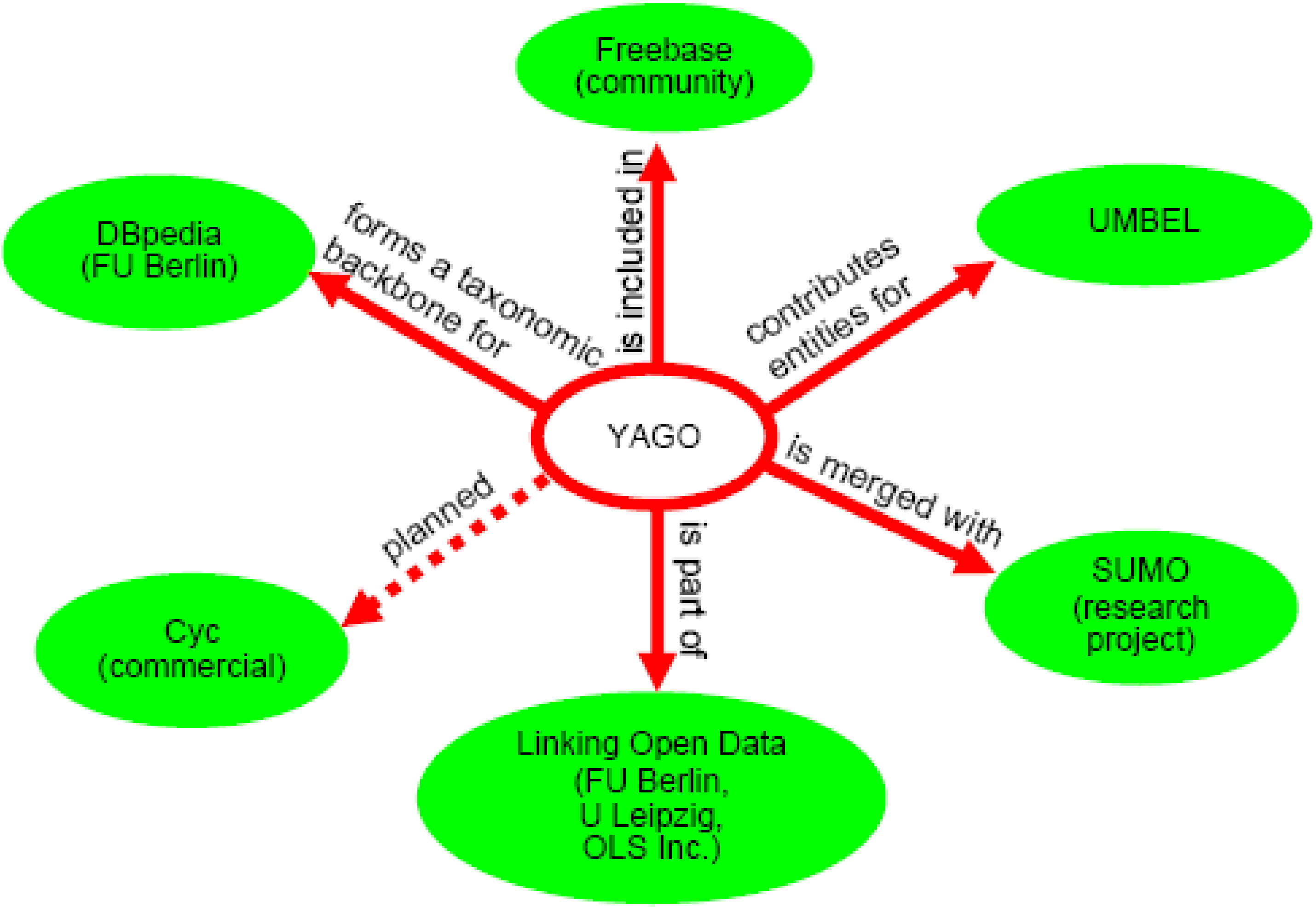


	Heuristic	#Eval	Precision
1	hasExpenses	46	100.0 % ± 0.0 %
2	hasInflation	25	100.0 % ± 0.0 %
3	hasLaborForce	43	97.67441% ± 0.0 %
4	during	232	97.48950% ± 1.838 %
5	ConceptualCategory	59	96.94342% ± 3.056 %
6	participatedIn	59	96.94342% ± 3.056 %
7	plays	59	96.94342% ± 3.056 %
8	establishedInYear	57	96.84294% ± 3.157 %
9	createdOn	57	96.84294% ± 3.157 %
10	originatesFrom	57	96.84294% ± 3.157 %
	...		
72	WordNetLinker	56	95.11911% ± 4.564 %
	...		
74	InfoboxType	76	95.08927% ± 4.186 %
75	hasSuccessor	53	94.86150% ± 4.804 %
	...		

Comparison with other ontologies



Applications



Questions?

Thank You

References

- YAGO: Yet Another Great Ontology, PhD Defense, Fabian M. Suchanek, Max-Planck Institute for Informatics, Saarbrücken