Knowledge Resources

Stefan Junghans
Mitja Richter

June 11 2014

Table of Contents

DIK

- Data
  Pure symbols, signals etc.
- Information
  Data with added meaning
- Knowledge
  The connection of, a collection of, or the rule-based inference from information

Knowledge Base

Knowledge is important for several scenarios like expert systems, web search, recommenders, manuals, etc. Merging/connecting knowledge leads to even more complex opportunities.
$\Rightarrow$ We need to store our knowledge somewhere!

- A database used for knowledge sharing and management
- Promotes the collection, organization and retrieval of knowledge

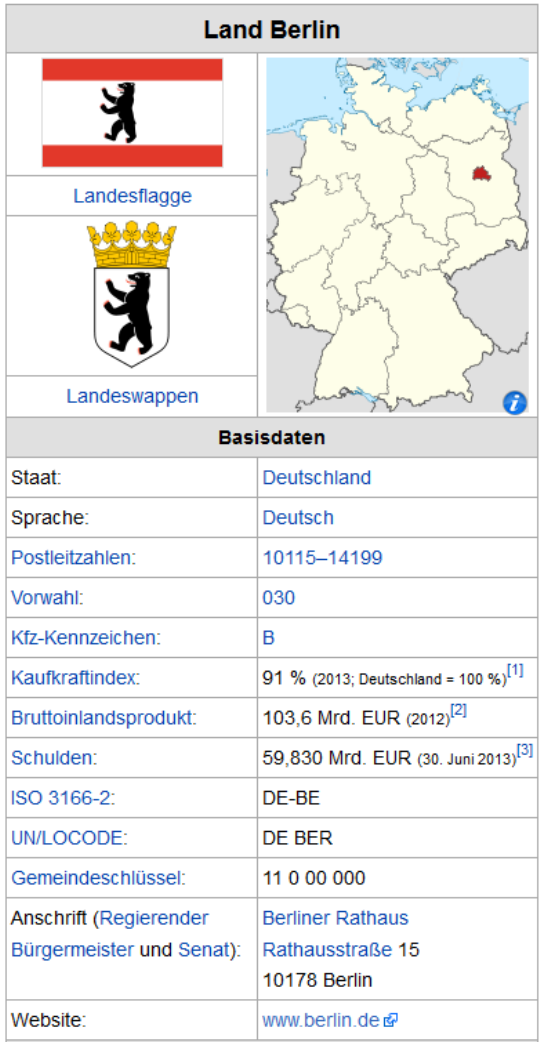
- Knowledge-based system
  - Reasoner
  - Knowledge Base

Definition

- Human-readable
  - Documents, Manuals, FAQs, ...
  - Wikipedia, transfermarkt.de, ...
- Machine-readable
  - System-readable forms
  - Ontologies
  - less interactive than human-readable forms
  - essential to the semantic web
  - DBpedia, Freebase, ...

DBpedia

- Founded by FU Berlin, Leipzig University and OpenLink Software in 2007
- DBpedia gathers structured information from Wikipedia
  - infoboxes
  - tables, lists
  - images
  - categories
  - links
  - geocoordinates

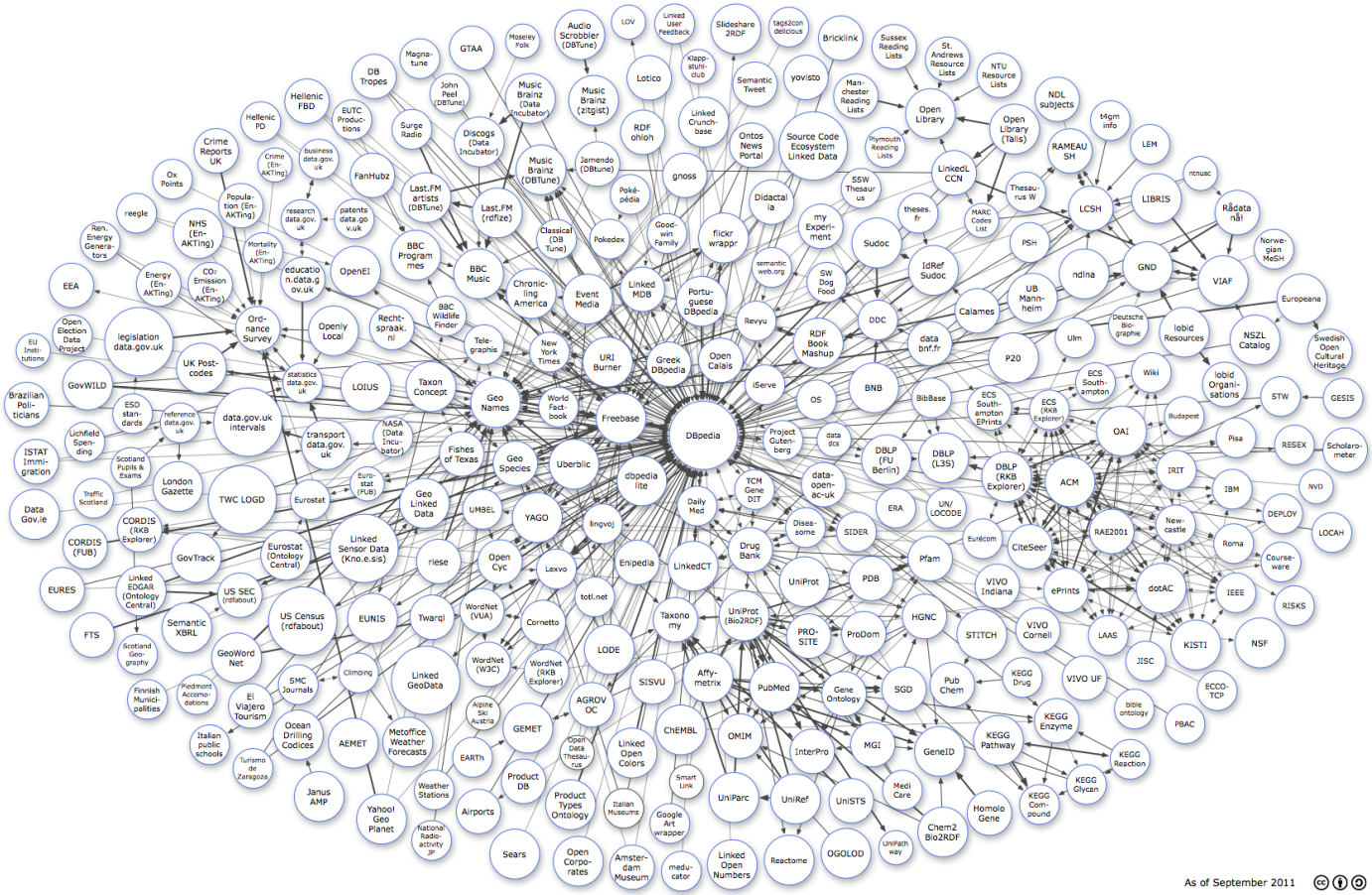


Land Berlin	
Landesflagge	
Landeswappen	
Basisdaten	
Staat:	Deutschland
Sprache:	Deutsch
Postleitzahlen:	10115–14199
Vorwahl:	030
Kfz-Kennzeichen:	B
Kaufkraftindex:	91 % (2013; Deutschland = 100 %)[1]
Bruttoinlandsprodukt:	103,6 Mrd. EUR (2012)[2]
Schulden:	59,830 Mrd. EUR (30. Juni 2013)[3]
ISO 3166-2:	DE-BE
UN/LOCODE:	DE BER
Gemeindeschlüssel:	11 0 00 000
Anschrift (Regierender Bürgermeister und Senat):	Berliner Rathaus Rathausstraße 15 10178 Berlin
Website:	www.berlin.de

Size

- english version currently describes 4 million things
  - 832.000 persons, 639.000 places, 372.000 creative works, 209.000 organizations, 226.000 species, 5.600 diseases
- 119 localized versions, in total:
  - 12.6 million unique things
  - 24.6 million links to images
  - 27.6 million links to external web pages
  - 45.0 million external links into other RDF datasets
  - 67.0 million links to Wikipedia categories
  - 41.2 million YAGO categories
- Linked Data!

Linked Open Data Cloud



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. `http://lod-cloud.net/` (9/2011)

Usage

- Use cases
  - sophisticated queries against Wikipedia
  - use data for web pages
  - “geographic applications” (connected to other geographic services like Geonames, CIA World Factbook, . . .)
  - annotate web content (multi-domain ontology)
  - many links to other datasets and vice versa
- Already used at
  - BBC
  - Amazon Datasets

Access and Representation

- De-referencable URIs in the form of `http://dbpedia.org/resource/[Name]`
- Information is represented in RDF format (or CSV)
- Provides a SPARQL endpoint
- Classifications:
  - Wikipedia Categorization
  - YAGO Classification
  - WordNet Synset Links
- Links to official homepages
- Owl:sameAs links (to DBpedia and external documents)
- Properties (rdfs, owl, foaf, dc, . . .)

Application - SNORQL

- Online SPARQL Explorer for DBpedia
- `http://dbpedia.org/snorql/`
- Example ...

Application - DBpedia Spotlight

- Transforms any text into DBpedia-annotated text
- Web Application
  `http://dbpedia-spotlight.github.io/demo/`
- REST Web Service
  `http://spotlight.dbpedia.org/rest/spot`
- Result can be returned in XML, JSON, HTML, RDFa or NIF
- Example ...

Freebase

- Initiated by Metaweb in 2007 (Google since 2010)
- User for Google Web/Knowledgegraph
- Graph based knowledge base
- Information is supplied directly by users or automatically through specific data pipelines (Wikipedia and Netflix)

Wladimir Wladimirowitsch Putin

Präsident der Russischen Föderation

Wladimir Wladimirowitsch Putin ist ein russischer Politiker. Er ist seit dem 7. Mai 2012 Präsident der Russischen Föderation; dasselbe Amt hatte er bereits von 2000 bis 2008 inne. Wikipedia

Geboren: 7. Oktober 1952 (Alter 61), Sankt Petersburg, Russland

Größe: 1,70 m

Ehepartnerin: Ljudmila Alexandrowna Putina (verh. 1983–2014)

Partei: Einiges Russland

Kinder: Mariya Putina, Yekaterina Putina

Eltern: Vladimir Spiridonovich Putin, Maria Ivanovna Shelomova

Wird auch oft gesucht Über 15 weitere ansehen

Ljudmila Alexandro... Putina
Exfrau

Barack Obama

Alina Maratowna Kabajewa

Dmitri Anatoljewi... Medwedew

Wiktor Janukowy...

Concepts

- Topics = nodes of the graph (Bob Dylan, Mercedes, . . .)
- Properties = edges of the graph (born in, produces, . . .)
- Type (songwriter, actor, car, . . .)
  - Compound Value Types
- Domain (music, business, . . .)
- Hierarchical URIs, e.g.
  `www.freebase.com/automotive/engine/horsepower`

Access

Read

- RDF API
  - `http://rdf.freebase.com/ns/en.al_gore`
- Search API (text search, ordered results after relevance)
  - `www.googleapis.com/freebase/v1/search?query=gore`
- Topic API (get the JSON for a specific topic)
  - `www.googleapis.com/freebase/v1/topic/en/al_gore`
- Data dumps
- Freebase Search Widget (jQuery plugin)

Access

Read/Write

- MQL API
- MQL Query Editor (`https://www.freebase.com/query`)
- Webpage (edit data, create views)

The Google APIs are available for Java, PHP, Python, .NET, JavaScript, Objective-C

Metaweb Query Language

- Read request sends and gets JSON
- `https://www.googleapis.com/freebase/v1/mqlread?query=_jsonInput_`

```
_jsonInput_:= {"type":"/music/artist",
               "name":"The Police",
               "album":[]}}
```

Metaweb Query Language

- Write request sends and gets JSON

- `https://www.googleapis.com/freebase/v1sandbox/mqlwrite?oauth_token=_token_&query=_query_`

```
_query_:= {"create":"unconditional",
           "id":null,
           "name":"Nowhere",
           "type":"/location/location"}
```

YAGO Yet Another Great Ontology

- YAGO (2008) YAGO2s (2012)
- developed at Max Planck Institute for Computer Science
- more than 10 million entities
- more than 447 million facts

Extraction from:

WordNet
A lexical database for English

UWN/MENTA — Universal Wordnet Project with MENTA extensions

WordNet Domains

GeoNames

Extracts from Wikipedia:

Title

Ronaldo

Category

Infobox

Fulltext

Relations:

- YAGO:
  about predefined relations with domain and range:

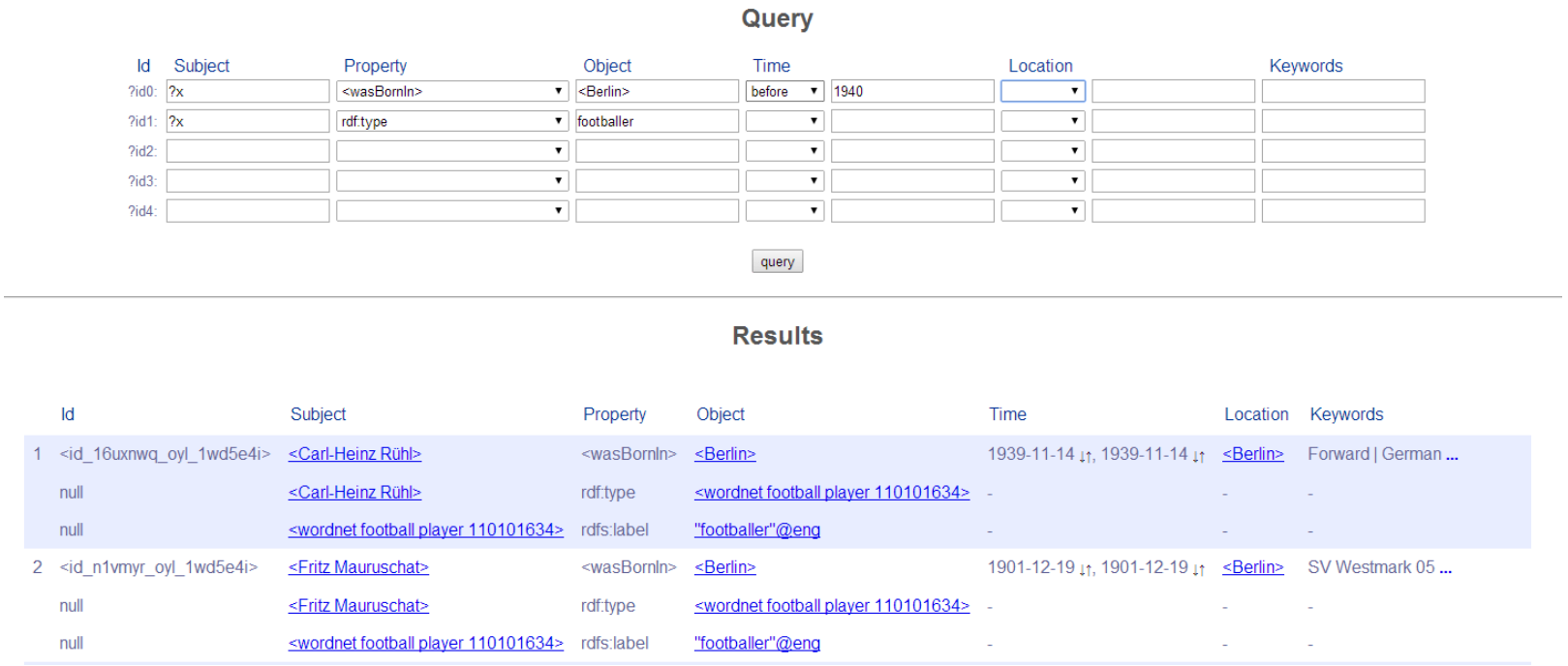
```
:hasSon  rdfs:domain  :Person;
                rdfs:range   :Man.
```

  adds temporal and spatial dimension to many entities/relations
- DBpedia:
  used to use words from infoboxes
  $\rightarrow$ length, length-in-km, length-km now uses also predefined relations

Relations: Time and Space

- many facts have a spatial and temporal dimension
  - Josef Ackermann is CEO of Deutsche Bank (2006-2012, London)
- not covered by rdf and SPARQL
- triplets to 5-lets
- SPARQL to SPOTL(X)
- still RDF

Demo spotlx



Query

Id	Subject	Property	Object	Time		Location		Keywords
?id0:	?x	<wasBornIn>	<Berlin>	before	1940			
?id1:	?x	rdf:type	footballer					
?id2:								
?id3:								
?id4:								

query

Results

	Id	Subject	Property	Object	Time	Location	Keywords
1	<id_16uxnwq_oyl_1wd5e4i>	<Carl-Heinz Rühl>	<wasBornIn>	<Berlin>	1939-11-14 ↓↑, 1939-11-14 ↓↑	<Berlin>	Forward | German ...
	null	<Carl-Heinz Rühl>	rdf:type	<wordnet football player 110101634>	-	-	-
	null	<wordnet football player 110101634>	rdfs:label	"footballer"@eng	-	-	-
2	<id_n1vmyr_oyl_1wd5e4i>	<Fritz Mauruschat>	<wasBornIn>	<Berlin>	1901-12-19 ↓↑, 1901-12-19 ↓↑	<Berlin>	SV Westmark 05 ...
	null	<Fritz Mauruschat>	rdf:type	<wordnet football player 110101634>	-	-	-
	null	<wordnet football player 110101634>	rdfs:label	"footballer"@eng	-	-	-

https://gate.d5.mpi-inf.mpg.de/webyagospotlx/WebInterface

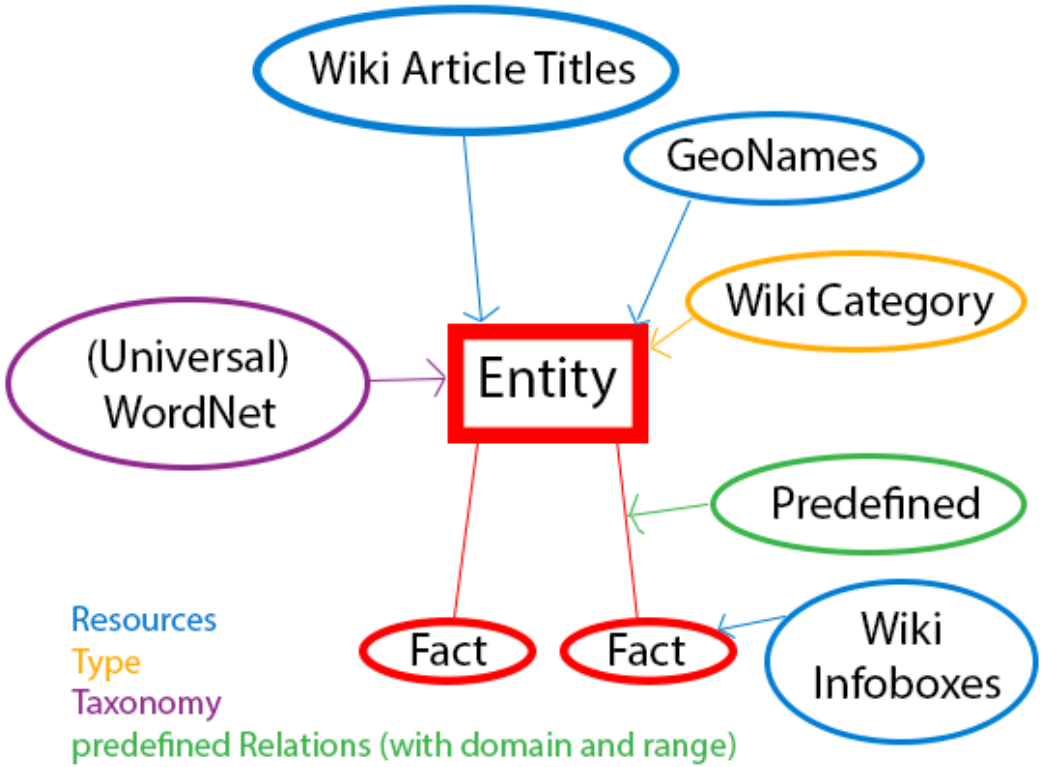
Precision and costs

- YAGO:
  - manually evaluated that 95
  - extractions process lasts 3 days
  - over 12 researcher

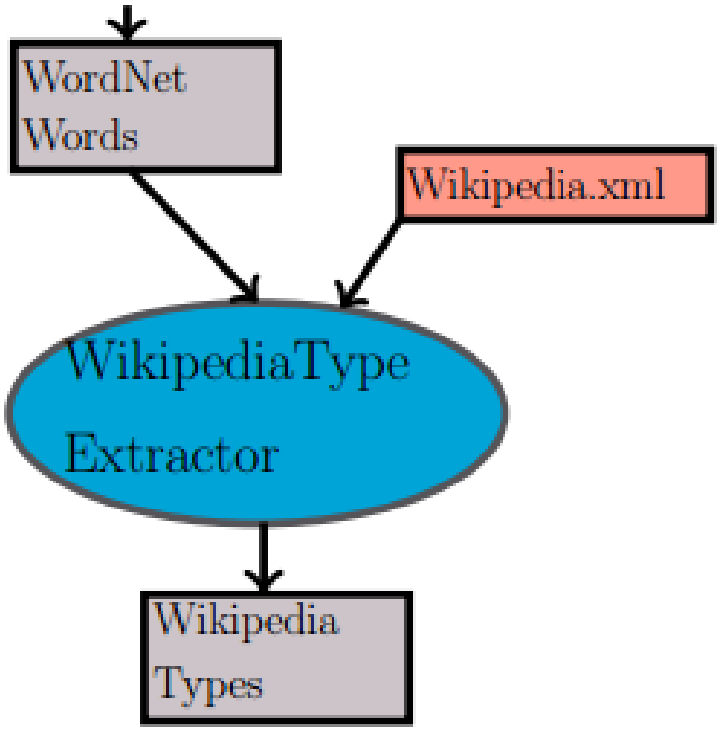
Why YAGO?

	YAGO	DBpedia
extracts:	title, category, infobox, fulltext	mainly infobox
relations:	predefined (domain, range)	taken from infobox
precision:	very good	not that good
costs:	very high	not that high

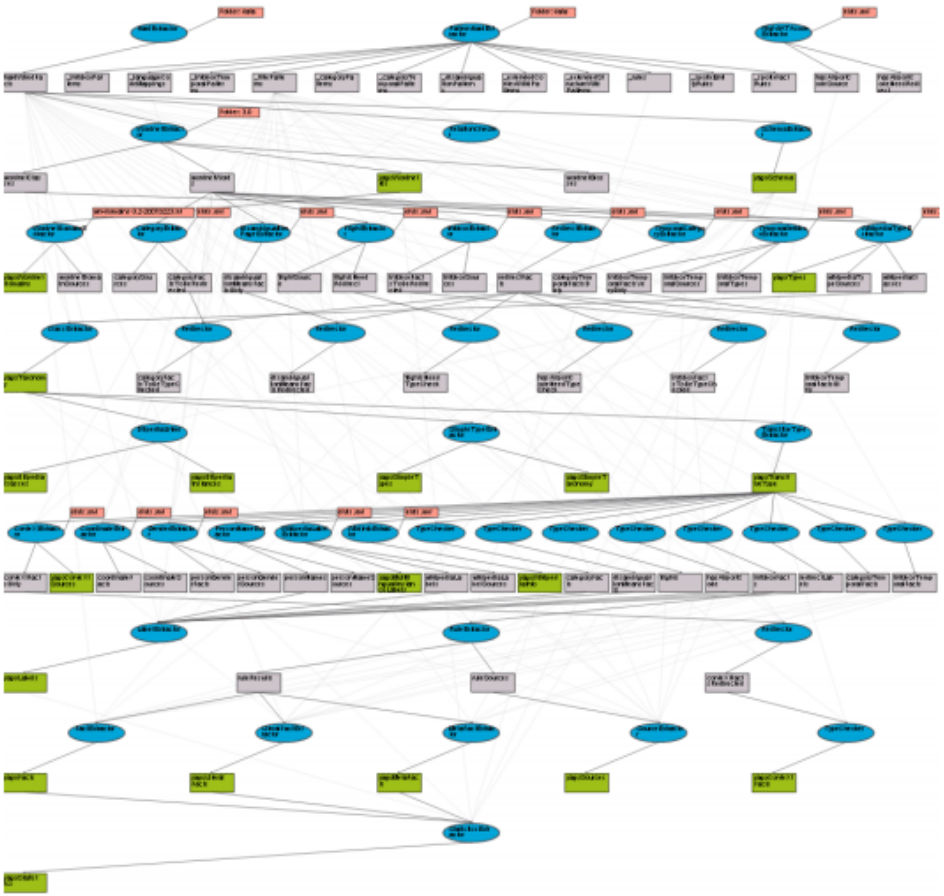
Extraction-process and Sources



YAGO2s

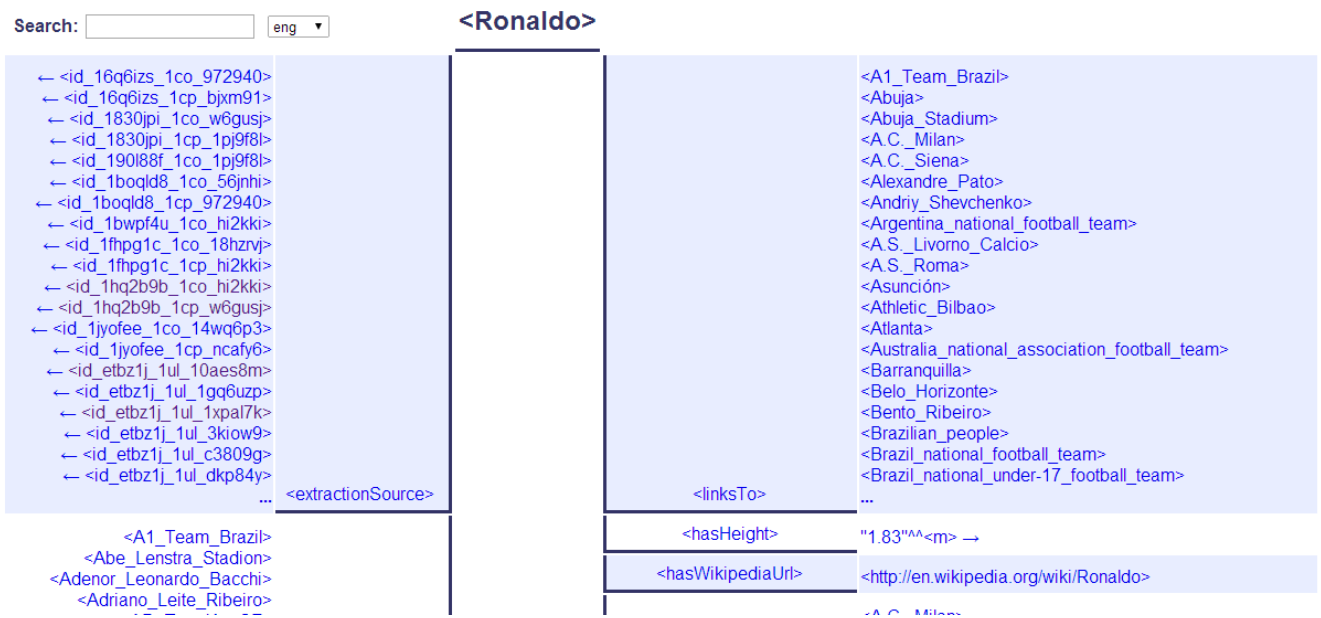

WordNet
Words
Wikipedia.xml
WikipediaType
Extractor
Wikipedia
Types

Visualization



http://resources.mpi-inf.mpg.de/yago-naga/yago/www2013demo/yago_demo_static/

Demo Ontology Browser



https://gate.d5.mpi-inf.mpg.de/webyagospotlx/Browser

When to use YAGO

- main goal
  near human-accuracy
- advantages
  - accuracy
  - coverage by GeoNames and WordNet
  - additional spatial and time domain
    space: 30 million, time 17 million
- disadvatages
  - hard to maintain/ not that up-to-date
  - not many and unreliable SPARQL/ SPOTL(x) endpoints

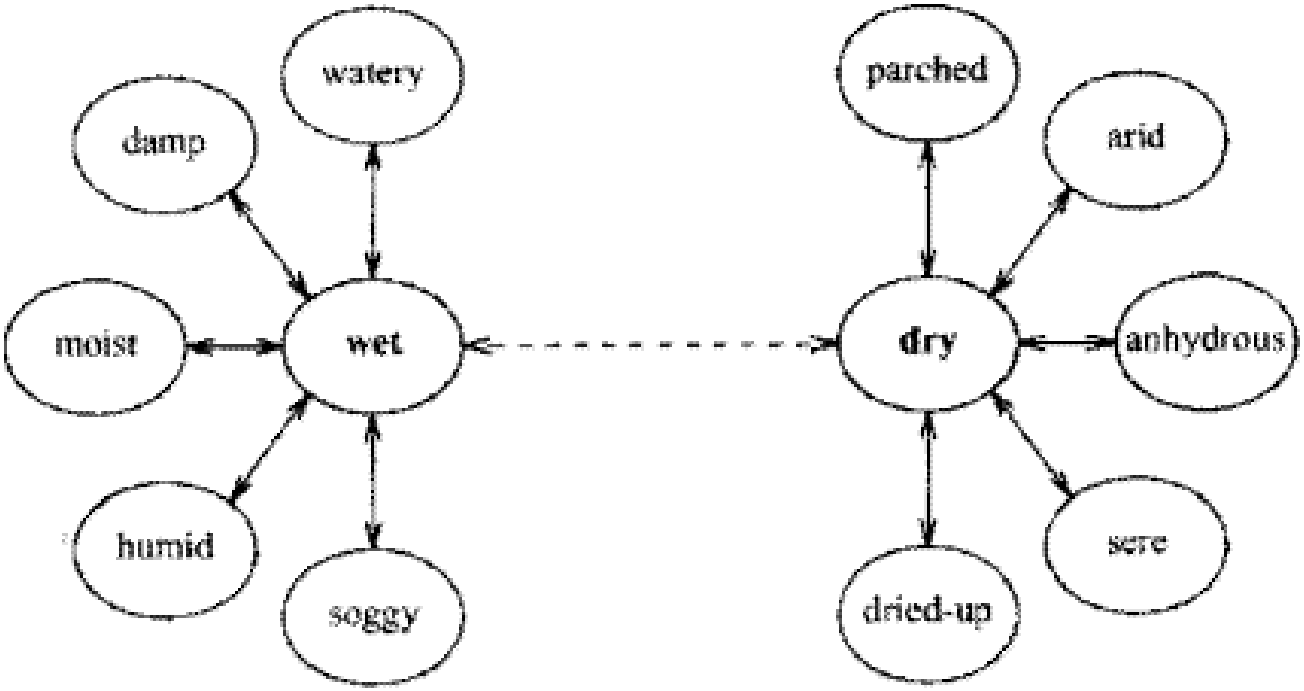
WordNet

WordNet
A lexical database for English

- development started 1985
- lexical database for English language
- covers most English nouns, verbs, adjectives, adverbs
- used in many applications (retrieval, translation)
- two kinds of semantic relations

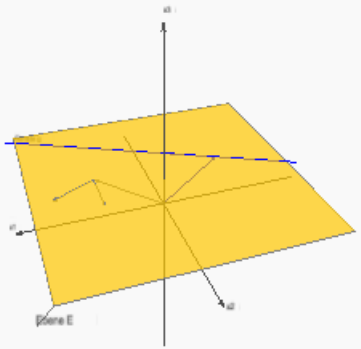
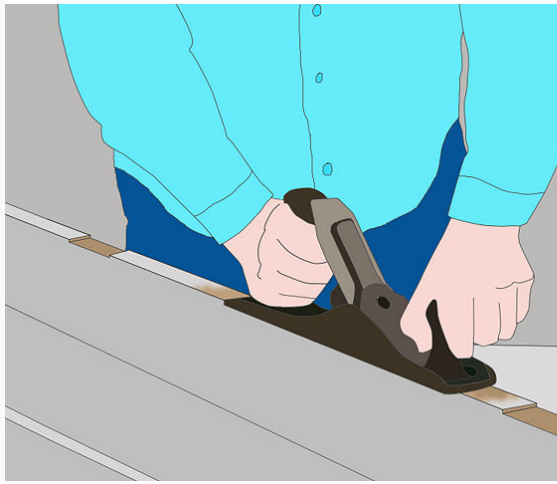
Synsets - Lexical (word-word) relation

- words are grouped into synonym sets - synsets
- sysnsets are the basic unit of meaning
- 120,000 synsets
- Synonym, Antonymy, Gradation



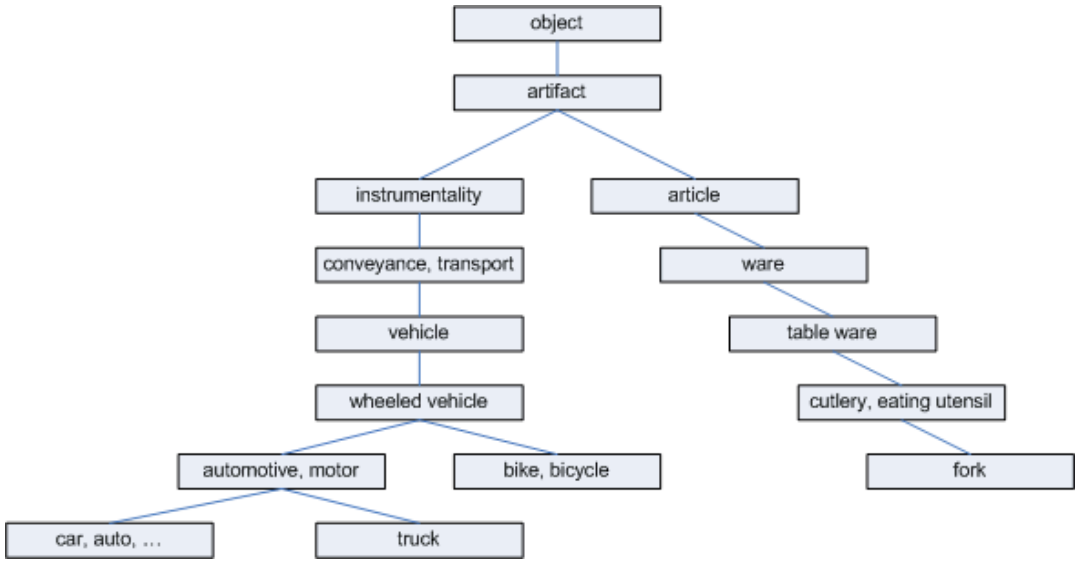
Homonyms

- Homonyms are represented in several Synsets

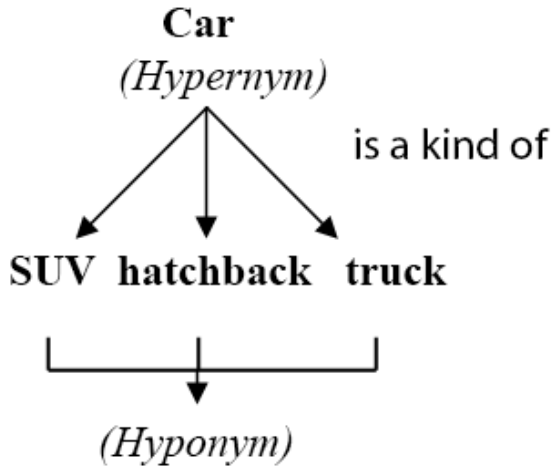
Linking between Synsets - Conceptual (concept-concept) relation

- Synsets are linked by
  - Hypernymy / Hyponymy
  - Meronymy / Holonymy
  - Entailment
  - Troponymy



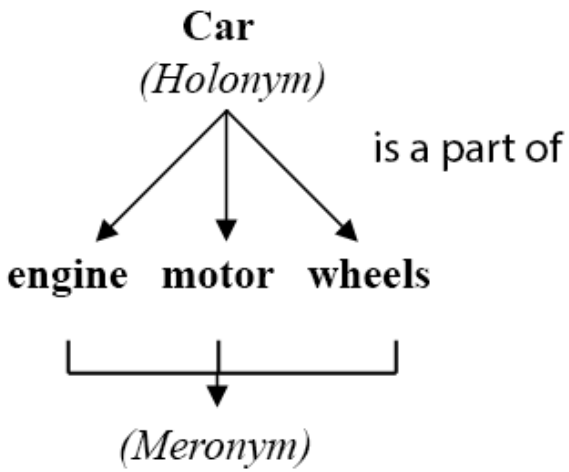
Hypernym / Hyponym

- A truck (Hyponym) is a kind of car (Hypernym)
- denotes more or less general concepts
- transitive
- experiments indicate knowledge about concepts is stored at superordinate (Hypernym) nodes and inherited downward



Meronym / Holonym (part/whole)

- The engine (Hyponym) is a part of a car (Hypernym)
- denotes more or less general concepts
- inheritance
- 3 kinds of meronomy
  - proper parts
  - substance
  - groups/members



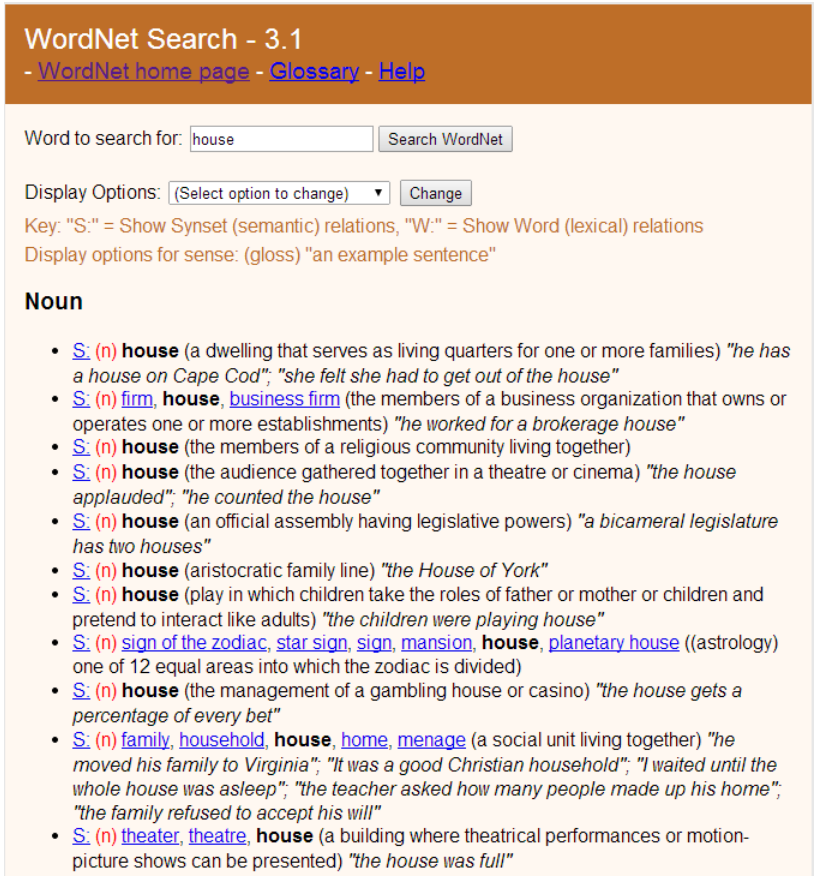
Semantic Relations between Verbs

- apart from Homonyms
  - Troponym:
    the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (lisp- talk)
  - Entailment:
    the verb Y is entailed by X if by doing X you must be doing Y (sleep - snore)

Problems and Limitations

- doesn't contain etymology, pronunciation or irregular verbs
- hard to modify or maintain
- no special domain vocabulary
- granularity

Demo



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: house Search WordNet

Display Options: (Select option to change) Change

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

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **house** (a dwelling that serves as living quarters for one or more families) *"he has a house on Cape Cod"; "she felt she had to get out of the house"*
- S: (n) firm, **house**, business firm (the members of a business organization that owns or operates one or more establishments) *"he worked for a brokerage house"*
- S: (n) **house** (the members of a religious community living together)
- S: (n) **house** (the audience gathered together in a theatre or cinema) *"the house applauded"; "he counted the house"*
- S: (n) **house** (an official assembly having legislative powers) *"a bicameral legislature has two houses"*
- S: (n) **house** (aristocratic family line) *"the House of York"*
- S: (n) **house** (play in which children take the roles of father or mother or children and pretend to interact like adults) *"the children were playing house"*
- S: (n) sign of the zodiac, star sign, sign, mansion, **house**, planetary house ((astrology) one of 12 equal areas into which the zodiac is divided)
- S: (n) **house** (the management of a gambling house or casino) *"the house gets a percentage of every bet"*
- S: (n) family, household, **house**, home, menage (a social unit living together) *"he moved his family to Virginia"; "It was a good Christian household"; "I waited until the whole house was asleep"; "the teacher asked how many people made up his home"; "the family refused to accept his will"*
- S: (n) theater, theatre, **house** (a building where theatrical performances or motion-picture shows can be presented) *"the house was full"*

http://wordnetweb.princeton.edu/perl/webwn

Demo - Python + TextBlob

Synsets

As you you WordNets basic Unit of meaning are Synset. These Synset contain a group of words with the same general meaning - Synonyms. So lets get some Synsets from WordNet.

In [2]:
```
import nltk
nltk.data.path.append('./nltk_data/')
from textblob import Word
word = Word("car")
word.synsets[:5]
```

Out[2]:
```
[Synset('car.n.01'),
 Synset('car.n.02'),
 Synset('car.n.03'),
 Synset('car.n.04'),
 Synset('cable_car.n.01')]
```

First we get all Synsets associated with the word car from textblob, by accessing the synsets property.

In [3]:
```
word.definitions[:5]
```

Out[3]:
```
['a motor vehicle with four wheels; usually propelled by an internal combustion engine',
 'a wheeled vehicle adapted to the rails of railroad',
 'the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant',
 'where passengers ride up and down',
 'a conveyance for passengers or freight on a cable railway']
```

To get the definitions of these synsets we are accessing the definitions property.

In [4]:
```
car = word.synsets[0]
```

Demo - Python + TextBlob

Today we are only interested in the kind of cars driving in the streets, so we assign this specific synset to a variable.

```
In [35]: car.lemma_names
```

```
Out[35]: ['car', 'auto', 'automobile', 'machine', 'motorcar']
```

The Synonyms in this Synset are called lemmas. We can get the string versions of them by accessing lemma_names

```
In [36]: car.hypernyms()
```

```
Out[36]: [Synset('motor_vehicle.n.01')]
```

Here we can see the hypernym / hyponym relation we have talked before. A car(hyponym) is kind of a motor vehicle(hypernym) and a minivan(hyponym) is kind of a car(hypernym)

```
In [38]: car.hyponyms()[:10]
```

```
Out[38]: [Synset('stanley_steamer.n.01'),
          Synset('hardtop.n.01'),
          Synset('loaner.n.02'),
          Synset('cruiser.n.01'),
          Synset('convertible.n.01'),
          Synset('minicar.n.01'),
          Synset('minivan.n.01'),
          Synset('hot_rod.n.01'),
          Synset('pace_car.n.01'),
          Synset('sports_car.n.01')]
```

Demo - Python + TextBlob

```
In [39]: car.member_holonyms()

Out[39]: []
```

Here we can see the same for meronyms/ holonyms. The car is not part of a bigger entity, so it has no holonym. On the other hand a car is a composition of many parts, so for example a car(holonym) has a hood(meronym).

```
In [40]: car.part_meronyms()

Out[40]: [Synset('gasoline_engine.n.01'),
          Synset('car_mirror.n.01'),
          Synset('third_gear.n.01'),
          Synset('hood.n.09'),
          Synset('automobile_engine.n.01'),
          Synset('grille.n.02'),
          Synset('automobile_horn.n.01'),
          Synset('rear_window.n.01'),
          Synset('car_window.n.01'),
          Synset('floorboard.n.02'),
          Synset('accelerator.n.01'),
          Synset('tail_fin.n.02'),
          Synset('window.n.02'),
          Synset('reverse.n.02'),
          Synset('glove_compartment.n.01'),
          Synset('first_gear.n.01'),
          Synset('buffer.n.06'),
          Synset('car_door.n.01'),
          Synset('roof.n.02'),
          Synset('auto_accessory.n.01'),
          Synset('car_seat.n.01'),
          Synset('high_gear.n.01'),
          Synset('fender.n.01'),
          Synset('stabilizer_bar.n.01'),
          Synset('bumper.n.02'),
          Synset('air_bag.n.01'),
          Synset('sunroof.n.01'),
          Synset('luggage_compartment.n.01'),
          Synset('running_board.n.01')]
```

Demo - Python + TextBlob

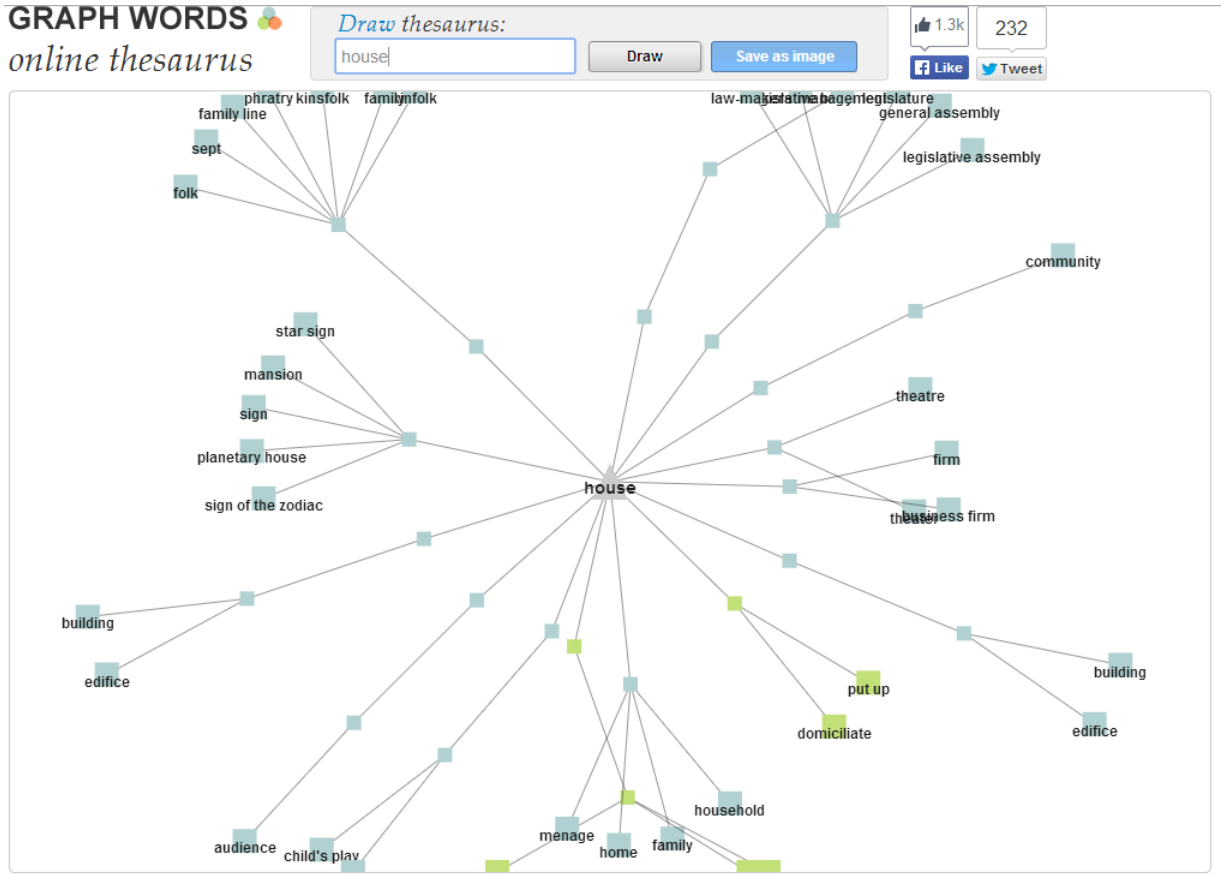
Semantic Similarity

Because Synsets are representated as a graph, we can measure the semantic similarity between these Synsets by path length between them. The scale is ranging from 1 (identical) to 0 (least similar). So lets compare some Synsets:

```
In [53]: from textblob.wordnet import Synset
         car = Synset("car.n.01")
         truck = Synset("truck.n.01")
         train = Synset("train.n.01")
         plane = Synset("airplane.n.01")
         pedestrian = Synset("pedestrian.n.01")

In [54]: car.path_similarity(car)

Out[54]: 1.0

In [55]: truck.path_similarity(car)

Out[55]: 0.3333333333333333

In [56]: train.path_similarity(car)

Out[56]: 0.125

In [57]: plane.path_similarity(car)

Out[57]: 0.1111111111111111

In [58]: pedestrian.path_similarity(car)

Out[58]: 0.07692307692307693
```

As expected the similarity decreases with less mutual properties, like wheels, engine etc.

Application



http://graphwords.com/

When to use Wordnet

Since WordNet is a lexical Database it has different UseCases:

- taxonomic backbone
- information retrieval
- translation
- Thesaurus

Other WordNets

There are many other WordNets for other languages

- Universal WordNet
  - 1.5 million words
  - 200 languages
  - based on WordNet
  - MENTA integration
    $\rightarrow$15 million words and names
- GermaNet
  - exclusive for German language
  - 93,000 synsets
  - 120,000 lexical units
  - for academics free
- EuroWordNet
  - Dutch, Italian, Spanish, German, French, Czech and Estonian
  - 200,000 synsets
  - not free
  - needs to be licensed

comparison

	DBpedia	Freebase
Sources	only Wikipedia	Wikipedia, Netflix, user input
+	has many outgoing links good categorization	many sources possibly more information
−	limited information	might seem a bit unstructured at times no SPARQL endpoint

comparison

	WordNet	YAGO
Sources	manual	Wikipedia, GeoNames, Word
+	good for word sense disambiguation really big	additional spatial and tempo information many sources very precise
−	too fine-grained no domain-specific vocabulary	not that up-to-date no good SPARQL endpoint