

DAT630

Entity Retrieval I.

01/11/2016

Krisztian Balog | University of Stavanger

Semantic Search

What is semantic search?

Google search results for "lisbon". The results include:

- Lisbon - Official Website - visitlisboa.com**: www.visitlisboa.com/TouristInfo
- Lisbon - Wikipedia, the free encyclopedia**: <https://en.wikipedia.org/wiki/Lisbon>
- Lisbon**: Aerial view of Lisbon with surrounding areas like Belém, Amadora, and Almada.
- Images for lisbon**: A grid of five images showing various landmarks in Lisbon.
- Lisbon, Portugal - Lonely Planet**: www.lonelyplanet.com/portugal/lisbon
- Lisbon Tourism: Best of Lisbon, Portugal - TripAdvisor**: www.tripadvisor.com...
- Lisbon**: Points of interest map showing locations like Belém Tower, Jerónimos Monastery, São Jorge Castle, etc.
- Points of interest**: Icons for Belém Tower, Jerónimos Monastery, São Jorge Castle, Lisbon Oceanarium, and Praça do Comércio.

Google search results for "lisbon things to see". The results include:

- Lisbon / Points of interest**: A grid of nine images of Lisbon landmarks.
- Things To Do In Lisbon - visitlisboa.com**: www.visitlisboa.com/ThingsToDo/
- Lisbon Thing To See - City Tours, Day Trips, and more**: www.visitor.com/lisbon
- Best Things to Do In Lisbon - holidaylettings.co.uk**: www.holidaylettings.co.uk/lisbon
- The Top 10 Things to Do in Lisbon - TripAdvisor - Lisbon**: www.tripadvisor.com/Attractions-g180064-Lisbon.html
- Best Things to Do in Lisbon | U.S.News Travel**: www.usnews.com/travel-best-trips/lisbon-best-things-to-do

Google search results for "kl1799". The results include:

- AMS → MUC**: Flight schedule from Amsterdam to Munich.
- KLM Flight 1799**: On-time - departs in 10 hours 50 mins.
- KLM Flight KL1799 - FlightAware**: flightaware.com/live/flight/KLM1799
- KL1799 Flight Status - FlightStats**: www.flightradar24.com/info/flights/KL1799
- KL1799 / KLM1799 - KLM Royal Dutch Airlines — Plane Fin...**: www.klm.com/klm1799
- KL1799 schedule (KLM flight Amsterdam -> Munich)**: <http://planeinfo.net/schedule/KL1799>

Google search results for "lenovo bios key". The results include:

- Bios Utility is entered by holding F2 while the ThinkPad logo is displayed after power on. Enter setup by pressing CTRL+Alt+F11 from a DOS prompt (you must be in DOS mode, not a DOS session under Windows). Press the F1 key and power the unit off. Hold the F1 key and power the unit on. Keep F1 held down until Easy setup appears.**
- How to access the BIOS - ThinkPad - Lenovo Support (HK)**: <https://support.lenovo.com/hk/en/documents/ht36045>
- How to enter Setup Utility (F1) or Boot Menu (F12) ... - Lenovo ..**: <https://support.lenovo.com/ie/Detail.page?LogSeq=00000000000000000000000000000000>
- How To Enter Bios Setup and Boot Menu On Lenovo G50 70 ...**: https://www.youtube.com/watch?v=_9_FgH0H4

What is semantic search?

- "Search with meaning"
- Improve search accuracy by **understanding** searcher intent and the **contextual meaning** of terms/documents/...
- Move beyond "ten blue links" (towards actually answering information needs) using rich context

Semantic search

- Centers around entities
 - "Who was the first human in outer space?"
 - "How tall is the Eiffel tower?"
 - "Who is Brad Pitt married to?"
 - "Where is the closest Starbucks?"
 - "Which airlines fly the Airbus A380?"
 - "What is the best Chinese restaurant in Montreal?"
- Entity/Attribute/Relationship retrieval
 - + social, + personal
 - + (hyper)local

Semantic search

- Combination of entity-related techniques, from various fields
 - Information Retrieval (IR)
 - Natural Language Processing (NLP)
 - Databases (DB)
 - Semantic Web (SW)

What is an entity?



What is an entity?

- Uniquely identifiable *thing* or *object*
 - "A thing with a distinct and independent existence"
- Characterized by having:
 - Unique ID
 - Name(s)
 - Type(s)
 - Attributes (/Descriptions)
 - Relationships to other entities

Entities...

- are meaningful units for organizing information
- are a key enabling component in semantic search

Entity Linking

Iranian POW negotiator holds talks with Iraqi ministers

The head of Iran's prisoner of war commission met with two Iraqi Cabinet ministers Saturday in a bid to glean information about thousands of Iranian POWs allegedly held in Iraq, the official Iraqi News Agency reported.

Iraq Foreign Minister **Mohammed Saed al-Sa'afat** told Abdullan al-Najafi that the two states needed to "speed up the closure of what remains from the POW and Missing-in-Action file," INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the 1980-88 **Iran-Iraq War**. The countries accuse each other of hiding POWs and preventing visits by the **International Committee of the Red Cross** to prisoner camps.

The ICRC representative in **Baghdad**, Manuel Bessler, told The Associated Press that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 since 1990.

More than 1 million people were held as civil law detainees in the largest exchange in southwest Asia (after Tehran).

[open in wikipedia](#)

Entity Retrieval

Google restaurants in montreal

About 66,300,000 results (0.29 seconds)

The Keg Steakhouse & Bar - Old Port
www.legategroup.com
4.3 ★★★★ 60 Google reviews | Google page
25 Rue Saint-Paul Est
Montreal, QC, Canada
+1 514-871-8050

Restaurant Kashima
www.kashimamontreal.com
4.3 ★★★★ 9 Google reviews
172 Avenue Greene
Westmount, QC, Canada
+1 514-834-0862

O.Noir
www.onoir.ca
4.3 ★★★★ 31 Google reviews
1611 Rue Sainte-Catherine Ouest
Montreal, QC, Canada
+1 514-837-9727

Restaurant Tran
phoquynh.com
4.3 ★★★★ 7 Google reviews
730 Rue Sainte-Catherine Ouest
Montreal, QC, Canada
+1 514-272-0992

L'Auberge Saint-Gabriel
www.lauberge-saint-gabriel.com
4.3 ★★★★ 12 Google reviews | Google page
426 St Gabriel St
Montreal, QC, Canada
+1 514-478-3301

Vago
www.restaurantvago.com
4.3 ★★★★ 10 Google reviews

Bistro - The Avenue
www.bistrotheavenue.com
3.7 ★★★★ 5 Google reviews
130 Avenue Greene
Westmount, QC, Canada
+1 514-839-6451

Map for restaurants in montreal

restaurants in Montreal - TripAdvisor
www.tripadvisor.com | Canada | Quebec | Montreal | TripAdvisor - Dining in Montreal, Quebec: See 65317 TripAdvisor traveler reviews of 4381 Montreal

Image taken from Milne and Witten (2008b). Learning to Link with Wikipedia. In CIKM '08.

Entity Linking/Retrieval

The screenshot shows a Google search results page for "Lisbon things to see". The top result is a snippet from "Things To Do In Lisbon - visitlisboa.com" featuring a grid of images of Lisbon landmarks like Belém Tower, Jerónimos Monastery, and São Jorge Castle. Below this are links to "Lisbon Things To See - City Tours, Day Trips, and more" (visitlisboa.com), "The Top 10 Things to Do in Lisbon - TripAdvisor" (tripadvisor.com), and "The Top 10 Things to Do in Lisbon - holidaylettings.co.uk" (holidaylettings.co.uk). A map of Lisbon is also visible on the right side of the results.

Meet the Data

Data collection

- Unstructured
 - Documents, web pages, snippets, ...
 - Semistructured
 - XML, RDF, ...
 - Structured
 - Relational DBs, RDF, ...
- } Often organized around entities

Single most popular semistructured data source

The screenshot shows the Wikipedia page for "New York City". The page title is "New York City" and it is described as "From Wikipedia, the free encyclopedia". The page content includes a brief history, a list of neighborhoods, and a sidebar with links to related topics like "New York" and "New York, New York". There are also sections for "Wikimedia Shop" and "Interaction". On the right side, there is a "Did you know" section with a box containing text about the city's population and a small image of the Statue of Liberty.

Knowledge Bases

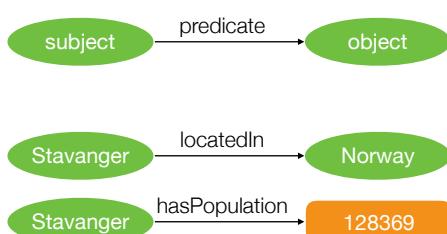
- Aimed at *machine understanding*
- Comprise a large set of assertions about the world
 - Describe (specific) entities and their relationships

RDF Data Model

- Resource Description Framework
 - Each resource is identified by a URI (Unique Resource Identifier)
 - (Entities = resources)
- Assertions are represented as triples
 - **Subject** (resource)
 - **Predicate** (relation)
 - **Object** (resource or literal)



RDF Data Model



Example

- How can this information be represented using RDF?

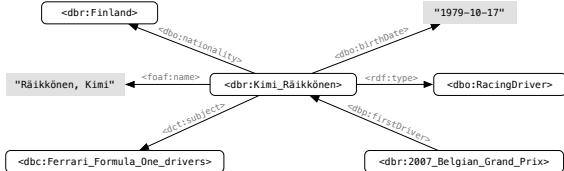
Kimi Räikkönen is a Finnish racing driver, born on October 17, 1979, currently driving for Ferrari in Formula One.

Example

Subject	Predicate	Object
<dbr:Kimi_Räikkönen>	<foaf:name>	"Räikkönen, Kimi"
<dbr:Kimi_Räikkönen>	<dbo:nationality>	<dbr:Finland>
<dbr:Kimi_Räikkönen>	<dbo:birthDate>	"1979-10-17"
<dbr:Kimi_Räikkönen>	<rdf:type>	<dbo:RacingDriver>
<dbr:Kimi_Räikkönen>	<dct:subject>	<dbc:Ferrari_Formula.One.drivers>

Knowledge bases

- Conceptually form a large, directed graph
- Also called **knowledge graphs** when the emphasis is on relationships between entities



SPARQL

- Structured query language for RDF

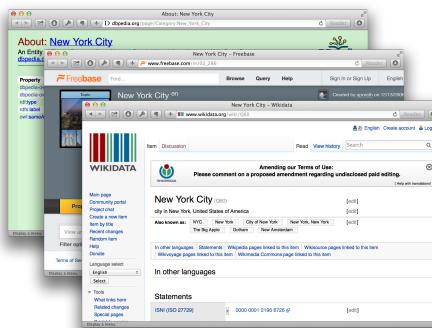
```
SELECT ?p WHERE {
?p has-profession Computer_Scientist .
?p has-gender Female .
?p occurs-with "semantic search"}
```

Early Attempt: Cyc

- Started in 1984 with the goal to manually build a knowledge base of everyday common knowledge
- ... still building and far from complete
- "one of the most controversial endeavors of the artificial intelligence history"

Popular (Public) Knowledge Bases

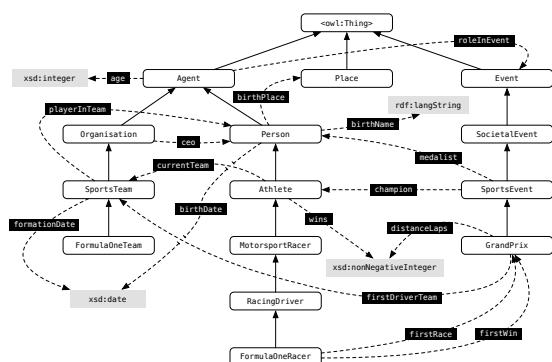
- DBpedia
- Freebase
- Wikidata



DBpedia

- "A database version of Wikipedia"
- Extracts RDF statements from Wikipedia articles
 - Mostly relies on infoboxes
 - Further homogenization or "normalization" is performed to achieve high data quality
 - Using manual mappings against the DBpedia Ontology

DBpedia Ontology



dbpedia.org



Towards a Public Data Infrastructure for a Large, Multilingual, Semantic Knowledge Graph



About: Kimi RäikkönenAn Entity of Type : **motorport_racer**, from Named Graph : <http://dbpedia.org/>, within Data Space : dbpedia.org

Kimi-Matias Räikkönen (Finnish pronunciation: [ˈkimi mætias ˈræik̊kœnɛn]; born 17 October 1979), nicknamed 'The Ice Man', is a Finnish racing driver currently driving for Ferrari in Formula One. After nine seasons racing in Formula One, in which during his first Ferrari stint he was the 2007 World Champion, he competed in the World Rally Championship in 2010 and 2011. In 2012, he returned to Formula One, driving for Lotus and continued to drive for Lotus in 2013. On 11 September 2013, Ferrari announced their signing of Räikkönen on a two-year contract, beginning in the 2014 season. His current contract with the team has been extended since and expires at the end of 2016. Räikkönen has also driven as a reserve driver for Williams in 2011. Having previously only raced in very junior open-wheel categories, he was given his Super Licence from the Fédération Internationale de l'Automobile (FIA) after a performance delivery promise by his team boss, Peter Sauber. He joined McLaren Mercedes in 2002, and became a title contender by finishing runner-up in the 2003 and 2005 Formula One World Championships. In 2007, he moved to Ferrari, where he won the 2007 World Drivers' Championship, beating McLaren drivers Lewis Hamilton and Fernando Alonso, respectively. Despite a 2008 25th place finish, he recovered by securing one more podium in the 2009 Formula One World Championship for Ferrari. In 2009, becoming the highest-paid driver in motor sport with an estimated wage of \$51 million per year. His move to Ferrari saw him secure his first Formula One World Drivers' Championship, beating McLaren drivers Lewis Hamilton and Fernando Alonso by one point, as well as becoming one of the very few drivers to win in their first season. In 2010, he equalised the record for fastest lap in Formula One history, and again got Fernando Alonso in trouble with a 2nd place. After a Sauber sponsorship, Räikkönen left the sport and joined Scuderia Ferrari F1 Team. In 2009, Räikkönen went into WRC to drive a Citroën C4 WRC for the Citroën Junior Team in the World Rally Championship for 2010. Along with relying, Räikkönen also competed in NASCAR, and finished third in the 2008 Busch Series. In 2010, he joined the Caterham Formula One Team. Räikkönen returned to F1 when he signed a 1-year deal with Lotus, along with the 2012 Formula one champion. In his debut season, he won the 2012 Abu Dhabi Grand Prix. His consistent performances allowed him to end the season 3rd in the Drivers' Championship. In 2008, Räikkönen was among the two Formula One drivers who made it into Forbes magazine's The Celebrity 100 list; the other being Fernando Alonso. He was 36th on the magazine's The Celebrity 100 list of 2008, and 41st the previous year. On the same list, as of 2008, he is listed as the 200th highest-paid celebrity overall and the 8th highest-paid sportsman behind Tiger Woods, David Beckham, Michael Jordan and Phil Mickelson. In 2009, Räikkönen was listed as the 2nd highest-paid athlete in the world, behind Woods. [\[more\]](#)

Property

Value

dbo:abstract

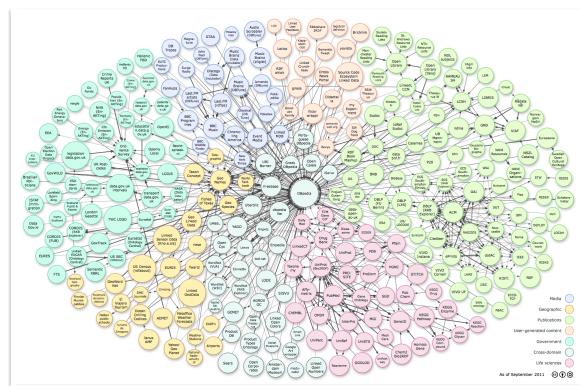
▪ Kimi-Matias Räikkönen (Finnish pronunciation: [ˈkimi mætias ˈræik̊kœnɛn]; born 17 October 1979), nicknamed 'The Ice Man', is a Finnish racing driver currently driving for Ferrari in Formula One. After nine seasons racing in Formula One, in which during his first Ferrari stint he was the 2007 World Champion, he competed in the World Rally Championship in 2010 and 2011. In 2012, he returned to Formula One, driving for Lotus and continued to drive for Lotus in 2013. On 11 September 2013, Ferrari announced their signing of Räikkönen on a two-year contract, beginning in the 2014 season. His current contract with the team has been extended since and expires at the end of 2016. Räikkönen has also driven as a reserve driver for Williams in 2011. Having previously only raced in very junior open-wheel categories, he was given his Super Licence from the Fédération Internationale de l'Automobile (FIA) after a performance delivery promise by his team boss, Peter Sauber. He joined McLaren Mercedes in 2002, and became a title contender by finishing runner-up in the 2003 and 2005 Formula One World Championships. In 2007, he moved to Ferrari, where he won the 2007 World Drivers' Championship, beating McLaren drivers Lewis Hamilton and Fernando Alonso, respectively. Despite a 2008 25th place finish, he recovered by securing one more podium in the 2009 Formula One World Championship for Ferrari. In 2009, becoming the highest-paid driver in motor sport with an estimated wage of \$51 million per year. His move to Ferrari saw him secure his first Formula One World Drivers' Championship, beating McLaren drivers Lewis Hamilton and Fernando Alonso by one point, as well as becoming one of the very few drivers to win in their first season. In 2010, he equalised the record for fastest lap in Formula One history, and again got Fernando Alonso in trouble with a 2nd place. After a Sauber sponsorship, Räikkönen left the sport and joined Scuderia Ferrari F1 Team. In 2009, Räikkönen went into WRC to drive a Citroën C4 WRC for the Citroën Junior Team in the World Rally Championship for 2010. Along with relying, Räikkönen also competed in NASCAR, and finished third in the 2008 Busch Series. In 2010, he joined the Caterham Formula One Team. Räikkönen returned to F1 when he signed a 1-year deal with Lotus, along with the 2012 Formula one champion. In his debut season, he won the 2012 Abu Dhabi Grand Prix. His consistent performances allowed him to end the season 3rd in the Drivers' Championship. In 2008, Räikkönen was among the two Formula One drivers who made it into Forbes magazine's The Celebrity 100 list; the other being Fernando Alonso. He was 36th on the magazine's The Celebrity 100 list of 2008, and 41st the previous year. On the same list, as of 2008, he is listed as the 200th highest-paid celebrity overall and the 8th highest-paid sportsman behind Tiger Woods, David Beckham, Michael Jordan and Phil Mickelson. In 2009, Räikkönen was listed as the 2nd highest-paid athlete in the world, behind Woods. [\[more\]](#)

dbo:birthDate

▪ 1979-10-17 (xsd:date)

dbo:birthPlace

▪ dbo:Espoo

Linking Open Data (LOD)**(re)Branding**

- Semantic Web data
- Linking Open Data
- Web of Data

Proprietary Knowledge Bases

Knowledge Graph



Entity Graph



Satori

... the knowledge graph is one of Google's biggest search milestones of the last decade...

— Amit Singhal, Google's director of search

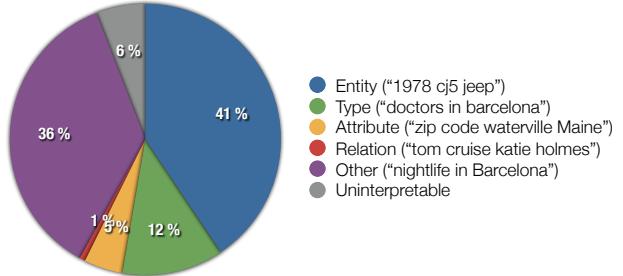
See: <https://www.youtube.com/watch?v=mmQI6VGvX-c>

RDFa, Microdata, ...

- Different protocols for marking up web pages
- schema.org
 - shared vocabulary
 - used by Google, Bing, Yandex, etc.
 - powers rich result snippets

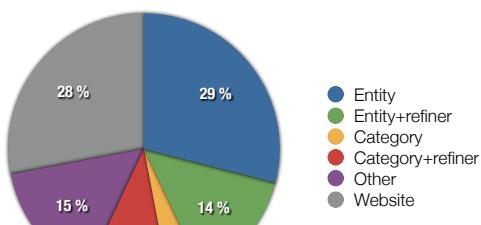
**Entity Retrieval****Entity retrieval**

Addressing information needs that are better answered by returning specific objects (entities) instead of just any type of documents.

Distribution of web search queries (Pound et al., 2010)

Pound, Mika, and Zaragoza (2010). Ad-hoc object retrieval in the web of data. In WWW '10.

Distribution of web search queries (Lin et al., 2011)



Lin, Pantel, Gamon, Kannan, and Fuxman (2012). Active objects. In WWW '12.

Entities

- Objects (or "things") with
 - Unique identifier
 - Name(s)
 - Attributes and/or description
 - Type(s)
 - Relationships to other entities

Ranking Entities with Ready-made Descriptions

"Entity homepages"



Document-based entity representations

- Most entities have a “home page”
- I.e., each entity is described by a document
- In this scenario, ranking entities is much like ranking documents
 - unstructured
 - semi-structured

Using Language Models

- Standard document retrieval methods applied on entity description documents
 - Just replacing d with e

$$P(e|q) \propto P(e)P(q|\theta_e) = \underbrace{P(e)}_{\text{Entity prior}} \prod_{t \in q} \underbrace{P(t|\theta_e)}_{\text{Entity language model}}^{n(t,q)}$$

Probability of the entity being relevant to any query
Multinomial probability distribution over the vocabulary of terms

Semi-structured entity representation

- Entity description documents are rarely unstructured
 - Different sections, fields, etc.

Audi A4
From Wikipedia, the free encyclopedia

The Audi A4 is a line of compact executive cars produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built in four generations and is based on Volkswagen's B platform. The first generation A4 succeeded the Audi 80. The automaker's internal numbering treats the A4 as a continuation of the Audi 80 lineage, with the initial A4 designated as the B5-series, followed by the B6, B7, and the current B8. The B8 A4 is built on the Volkswagen Group's MLB platform shared with many other Audi models and potentially one Porsche model within Volkswagen Group.^[2]

The Audi A4 automobile layout consists of a longitudinally oriented engine at the front, with transaxle-type transmissions mounted at the rear of the engine. The cars are front-wheel drive, or on some models, "quattro" all-wheel drive.

The A4 is available as a saloon/sedan and estate/wagon. The second (B6) and third generations (B7) of the A4 also had a convertible version, but the B8 version of the convertible became a variant of the Audi A5 instead Audi got back into the compact executive coupé segment. The Facebook fans of the Audi A4 page are more than 870,000.

Contents [show]

Article **Talk** **Create account** **Log in** **Read** **Edit** **View history** **Search**

How to rank entities in knowledge bases?

```

dbpedia:Audi_A4

foaf:name          Audi A4
rdfs:label         Audi A4
rdfs:comment       The Audi A4 is a compact executive car
                   produced since late 1994 by the German car
                   manufacturer Audi, a subsidiary of the
                   Volkswagen Group. The A4 has been built [...]
dbpprop:production 1994
                   2001
                   2005
                   2008
rdf:type            dbpedia-owl:MeanOfTransportation
dbpedia-owl:manufacturer dbpedia-owl:Automobile
dbpedia-owl:class    dbpedia:Audi
owl:sameAs          dbpedia:Compact_executive_car
is dbpedia-owl:predecessor of  freebase:Audi_A4
is dbpprop:similar of   dbpedia:Audi_A5
                                         dbpedia:Cadillac_BLS

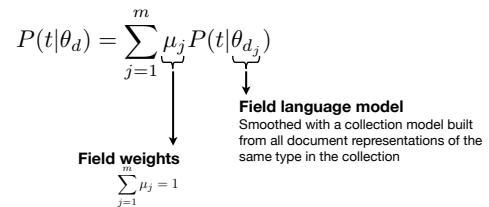
```

How to rank entities in knowledge bases?

- Represent entities as **fielded documents**

Fielded models

- Fielded extensions of document retrieval methods
- E.g., Mixture of Language Models (MLM)



Setting field weights

- Heuristically
 - Proportional to the length of text content in that field, to the field's individual performance, etc.
- Empirically (using training queries)
- Problems
 - Number of possible fields is huge
 - It is not possible to optimise their weights directly
- Entities are sparse w.r.t. different fields
 - Most entities have only a handful of predicates

Predicate folding

- **Idea:** reduce the number of fields by grouping them together
- Grouping based on
 - type
 - manually determined importance

Predicate Folding

Name	{foaf:name rdfs:label rdfs:comment}	Audi A4 Audi A4 The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]
Attributes	dbpprop:production	1994 2001 2005 2008
Out-relations	rdf:type	dbpedia-owl:MeanOfTransportation
	dbpedia-owl:manufacturer	dbpedia-owl:Automobile
	dbpedia-owl:class	dbpedia:Audi
In-relations	owl:sameAs	dbpedia:Compact_executive_car
	is dbpedia-owl:predecessor of	freebase:Audi_A4
	is dbpprop:similar of	dbpedia:Audi_A5
		dbpedia:Cadillac_BLS

Entity Resolution

- Need to replace entity URIs with their names
- so that they become "searchable" terms

Name	{foaf:name rdfs:label rdfs:comment}	Audi A4 Audi A4 The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]
Attributes	dbpprop:production	1994 2001 2005 2008
Out-relations	rdf:type	dbpedia-owl:MeanOfTransportation
	dbpedia-owl:manufacturer	dbpedia-owl:Automobile
	dbpedia-owl:class	dbpedia:Audi
In-relations	owl:sameAs	dbpedia:Compact_executive_car
	is dbpedia-owl:predecessor of	freebase:Audi_A4
	is dbpprop:similar of	dbpedia:Audi_A5
		dbpedia:Cadillac_BLS

Mean of transportation Audi A5