

Utilizing Knowledge Bases in Text-centric Information Retrieval

Laura Dietz (@lauradietz99)

University of New Hampshire

Alexander Kotov (@rusillini)

Wayne State University

Edgar Meij (@edgarmeij)

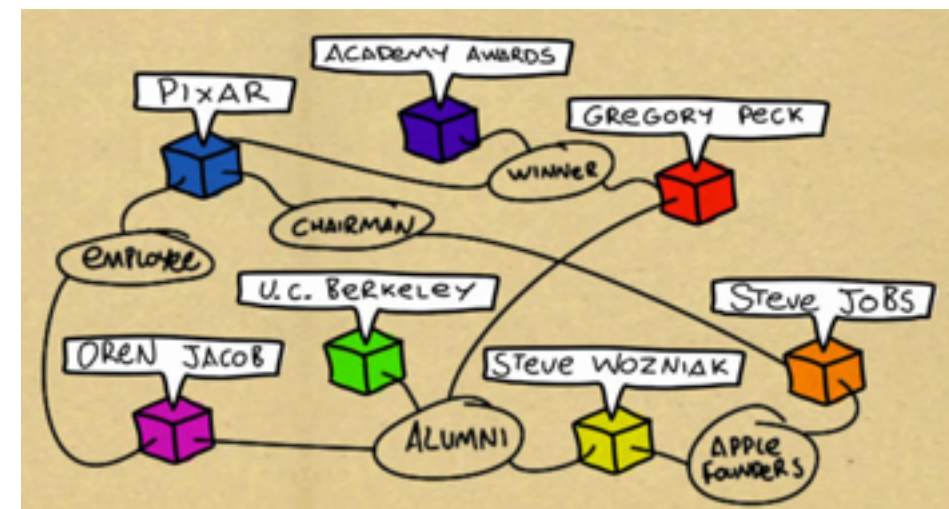
Bloomberg

Entity?

- Uniquely identifiable *thing* or *object*
 - “A thing with a distinct and independent existence”
 - people, places, products, companies, etc. etc.

What's so special about entities?

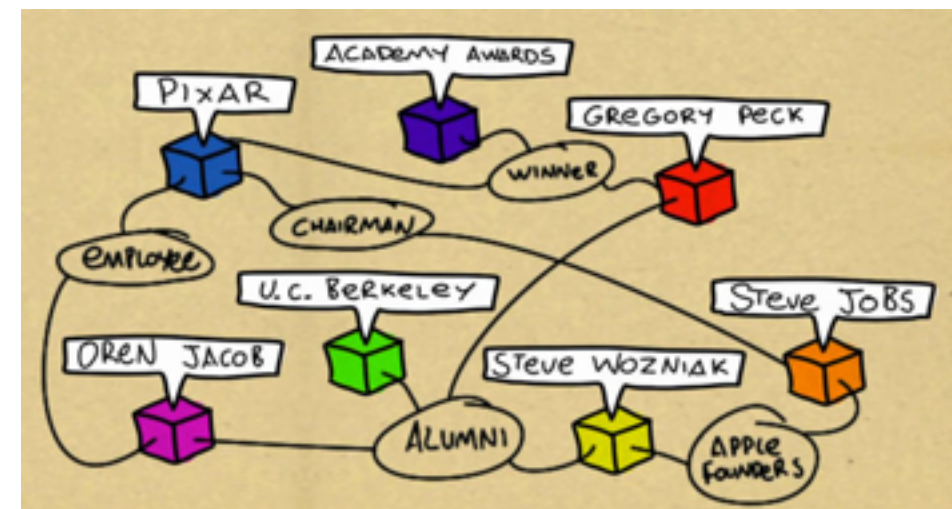
- ID
- Name(s)
- Type(s)
- Attributes (/Descriptions)
- Relationships to other entities



Knowledge graphs

- The “backbone” of semantic search
- They define
 - entities
 - attributes
 - types
 - relations
 - (provenance, sometimes)
 - and more
 - external links, homepages, features, ...

ICTIR 2016 Tutorial on Utilizing KGs in Text-centric IR



Knowledge graphs

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:Automobile dbpedia:Audi dbpedia:Compact_executive_car
dbpedia-owl:manufacturer	freebase:Audi A4
dbpedia-owl:class	dbpedia:Audi_A5
owl:sameAs	dbpedia:Cadillac_BLS
is dbpedia-owl:predecessor of	
is dbpprop:similar of	

Entity Linking/Retrieval

The screenshot shows a Google search for 'newark'. The search bar at the top contains 'newark' and the Google logo. Below the search bar, there are tabs for 'All', 'Maps', 'News', 'Images', 'Shopping', 'More', and 'Search tools'. The search results are displayed below the tabs. The first result is 'Electronic Components Distributor Newark element14' with the URL 'www.newark.com/'. Below this, there is a snippet of text: 'Same day shipping for even the smallest of orders, on a huge range of technology products from Newark element14. New items from leading brands added ...'. To the right of this result is a knowledge panel for 'Newark element14 Corporation'. The panel includes the company logo, a share icon, and a link to 'newark.com'. Below the link, there is a description: 'Newark element14, sometimes called Newark Corporation, Newark or Newark Element14, is a Chicago-based electronic components distribution company serving North America and parts of Central and South America. Wikipedia'. Below the description, there are fields for 'Headquarters: Chicago, IL', 'Founded: 1934', and 'Parent organization: Premier Farnell'. At the bottom of the panel, there are social media links for Facebook, LinkedIn, Twitter, and Google+. The second result is 'Newark, New Jersey - Wikipedia, the free encyclopedia' with the URL 'https://en.wikipedia.org/wiki/Newark,_New_Jersey'. Below this, there is a snippet of text: 'Newark is the largest city (by population) in the U.S. state of New Jersey, and the county seat of Essex County. One of the nation's major air, shipping, and rail ...'. Below the snippet, there are fields for 'State: New Jersey', 'Area code(s): 862/973', 'County: Essex', and 'Area rank: 103rd of 566 in state; 1st of 22 in co...'. The third result is 'Newark - Wikipedia, the free encyclopedia' with the URL 'https://en.wikipedia.org/wiki/Newark'. Below this, there is a snippet of text: 'Newark commonly refers to. Newark, New Jersey, United States. Newark Liberty International Airport, New Jersey. Newark-on-Trent, Nottinghamshire, England.'. At the bottom of the search results, there is a section titled 'In the news' with a small image of a police car and the headline '18-year-old is 2nd teen to die in Newark shootings this Labor Day weekend'.

Google

newark

www.google.com/#q=newark

Sign in

All Maps News Images Shopping More Search tools

About 99,800,000 results (0.59 seconds)

Electronic Components Distributor Newark element14
www.newark.com/
Same day shipping for even the smallest of orders, on a huge range of technology products from Newark element14. New items from leading brands added ...

Results from newark.com


Products
at Newark element14. Competitive prices from the leading distributor.

Contact Us
Contact Us. We're here to make your life easier. How can we ...


Newark, New Jersey - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Newark,_New_Jersey Wikipedia
Newark is the largest city (by population) in the U.S. state of New Jersey, and the county seat of Essex County. One of the nation's major air, shipping, and rail ...
State: **New Jersey** Area code(s): **862/973**
County: **Essex** Area rank: 103rd of 566 in state; 1st of 22 in co...


Newark - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Newark> Wikipedia
Newark commonly refers to. Newark, New Jersey, United States. Newark Liberty International Airport, New Jersey. Newark-on-Trent, Nottinghamshire, England.

In the news

 **18-year-old is 2nd teen to die in Newark shootings this Labor Day weekend**

Newark element14 Corporation







 newark.com

Newark element14, sometimes called Newark Corporation, Newark or Newark Element14, is a Chicago-based electronic components distribution company serving North America and parts of Central and South America. [Wikipedia](#)

Headquarters: [Chicago, IL](#)
Founded: 1934
Parent organization: [Premier Farnell](#)

Profiles

 Facebook  LinkedIn  Twitter  Google+

People also search for [View 15+ more](#)

Entity Linking/Retrieval

www.google.com/#q=newark+usa

Google

newark usa

Sign In

All Maps Images News Shopping More Search tools

About 35,500,000 results (0.36 seconds)

Electronic Components Distributor Newark element14
www.newark.com/
Same day shipping for even the smallest of orders, on a huge range of technology products from Newark element14. New items from leading brands added ...
[Products](#) · [Contact Us](#) · [Test and Measurement](#) · [About Us](#)



Newark, New Jersey - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Newark,_New_Jersey · Wikipedia
Newark is the largest city (by population) in the U.S. state of New Jersey, and the county seat of Essex County. One of the nation's major air, shipping, and rail ...
[Essex County, New Jersey](#) · [Mayors of Newark](#) · [List of neighborhoods in ...](#) · [Ivy Hill](#)

Newark - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Newark> · Wikipedia
Newark commonly refers to. Newark, New Jersey, United States. Newark Liberty International ... Newark, Delaware, USA. Newark ... Newark, New Jersey, USA.

Newark USA
newarkusa.blogspot.com/
A fotojournal about LIVING in Newark USA, New Jersey's largest and most ... by the author of the foto-essay website RESURGENCE CITY: Newark USA. Classic.

City of Newark, New Jersey
www.ci.newark.nj.us/ · Newark
Official City of Newark website with department information, municipal contacts, City calendar and more.

Current Local Time in Newark, New Jersey, USA - Timeanddate.com

Newark
City in New Jersey

Newark is the largest city in the U.S. state of New Jersey, and the county seat of Essex County. One of the nation's major air, shipping, and rail hubs, the city had a population of 277,140 in 2010, ... [Wikipedia](#)

Weather: 71°F (22°C), Wind N at 14 mph (23 km/h), 49% Humidity


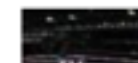


Hotels: 3-star averaging \$149. [View hotels](#)

Local time: Monday 10:10 AM

Population: 278,427 (2013)

Mayor: [Ras Baraka](#)

Points of interest [View 10+ more](#)

Entity Linking/Retrieval

www.google.com/#q=newark+delaware+usa

Google

newark delaware usa

Sign In

All Maps Images News Shopping More Search tools

About 1,620,000 results (0.48 seconds)

Newark, Delaware - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Newark,_Delaware - Wikipedia
Newark is a city in New Castle County, Delaware, 12 miles (19 km) west- southwest of Wilmington. According to the 2010 Census, the population of the city is ...
[History](#) · [Geography](#) · [Demographics](#) · [Education](#)

University of Delaware
www.udel.edu/ - University of Delaware
The University of Delaware is a diverse institution of higher learning, fostering ... [Contact Us](#); [University of Delaware Newark, DE 19716 USA](#); P: 302-831-2792 ...

Best Places to Live in Newark, Delaware - Sperling's Best Places
www.bestplaces.net/city/delaware/newark
Newark, Delaware Map. 31,655 Up 6.6% Population; 30.6% ... [Real Estate in Newark](#) [Comparison](#).
[Compare Newark, Delaware to any other place in the USA](#).

CareersUSA Delaware - CareersUSA Putting people to work
careersusa.com/Locations/UnitedStates/Delaware.aspx
1450 Capitol Trail, Suite 111. Newark, DE 19711. TEL: (302) 737-3600 - FAX: (302) 737-3606.
newark@careersusa.com · [CareersUSA Wilmington](#). [National](#) ...

Images for newark delaware usa [Report images](#)

Newark
City in Delaware

Newark is a city in New Castle County, Delaware, 12 miles west-southwest of Wilmington. According to the 2010 Census, the population of the city is 31,454. Newark is the home of the University of Delaware. [Wikipedia](#)

Area: 8.88 mi²
Weather: 73°F (23°C), Wind NW at 8 mph (13 km/h), 55% Humidity
Hotels: 3-star averaging \$129. [View hotels](#)
Local time: Monday 10:11 AM
Population: 32,549 (2013)
Area code: 302
Unemployment rate: 4.1% (Apr 2015)

Colleges and Universities [View 14 more](#)

Entity Retrieval

www.google.com/#q=restaurants+in+newark+delaware+usa


Google

restaurants in newark delaware usa

Sign In

All Maps Shopping Images News More Search tools


About 825,000 results (0.73 seconds)




Rating Cuisine Price Hours

Labor Day might affect these hours


Iron Hill Brewery & Restaurant
4.2 ★★★★★ (113) · \$\$ · American
Microbrews & American comfort fare
147 E Main St
Opens at 11:30 AM



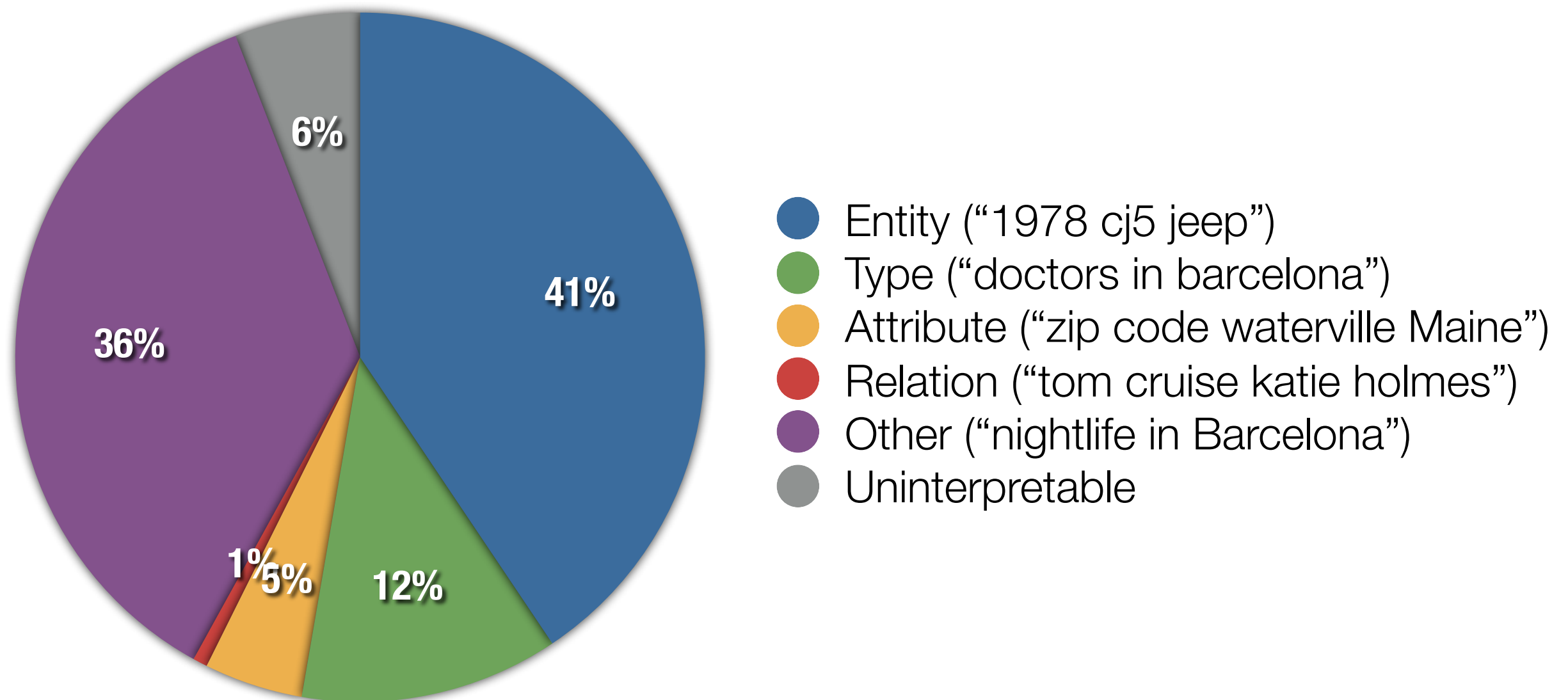
Caffè Gelato Restaurant
4.0 ★★★★★ (119) · \$\$ · Italian
Upscale Italian-American fare & desserts
90 E Main St
Opens at 11:00 AM



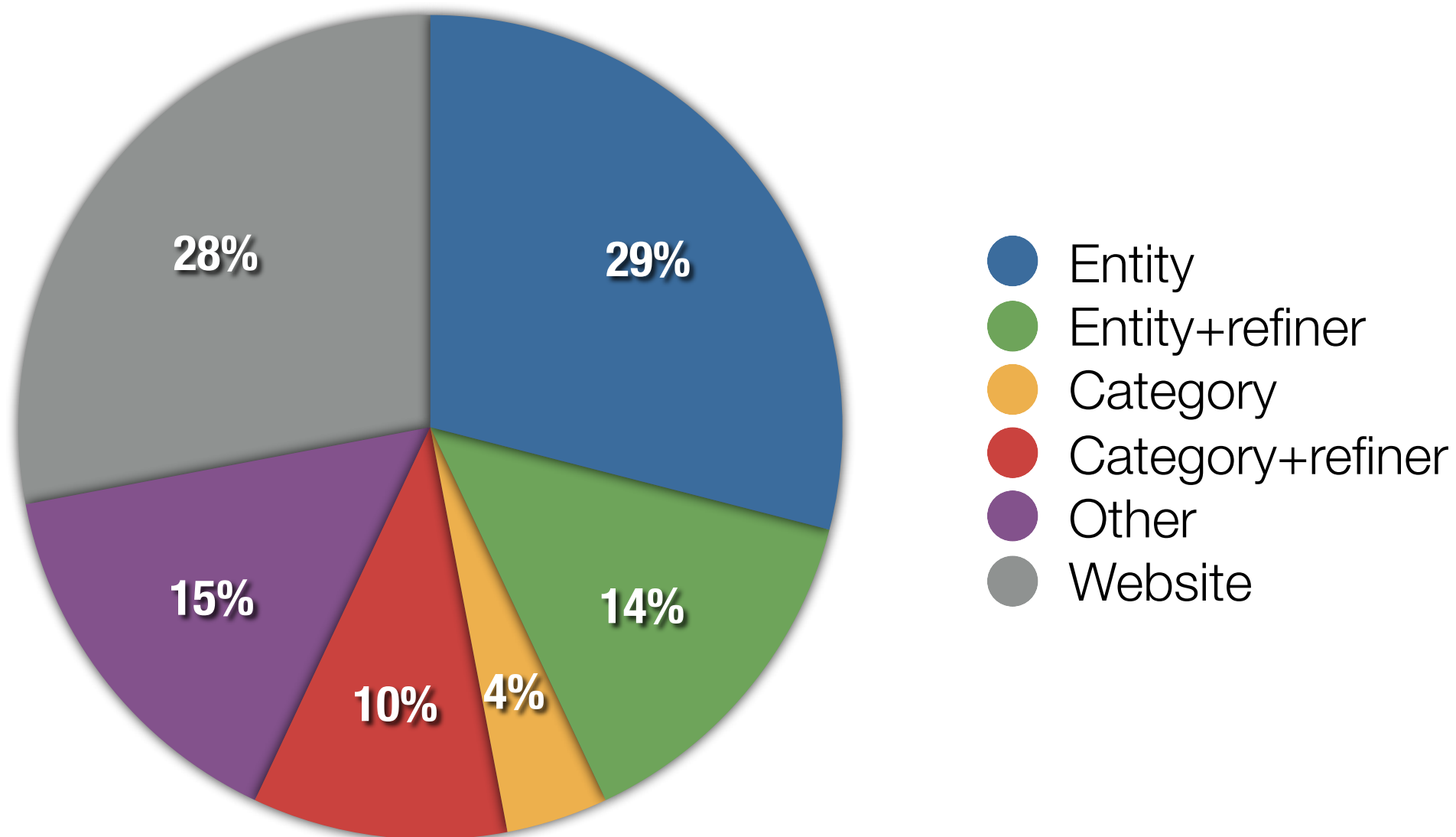
Taverna
4.6 ★★★★★ (43) · Italian
Warm, cozy Italian atmosphere & drinks



Distribution of web search queries [Pound et al. 2010]



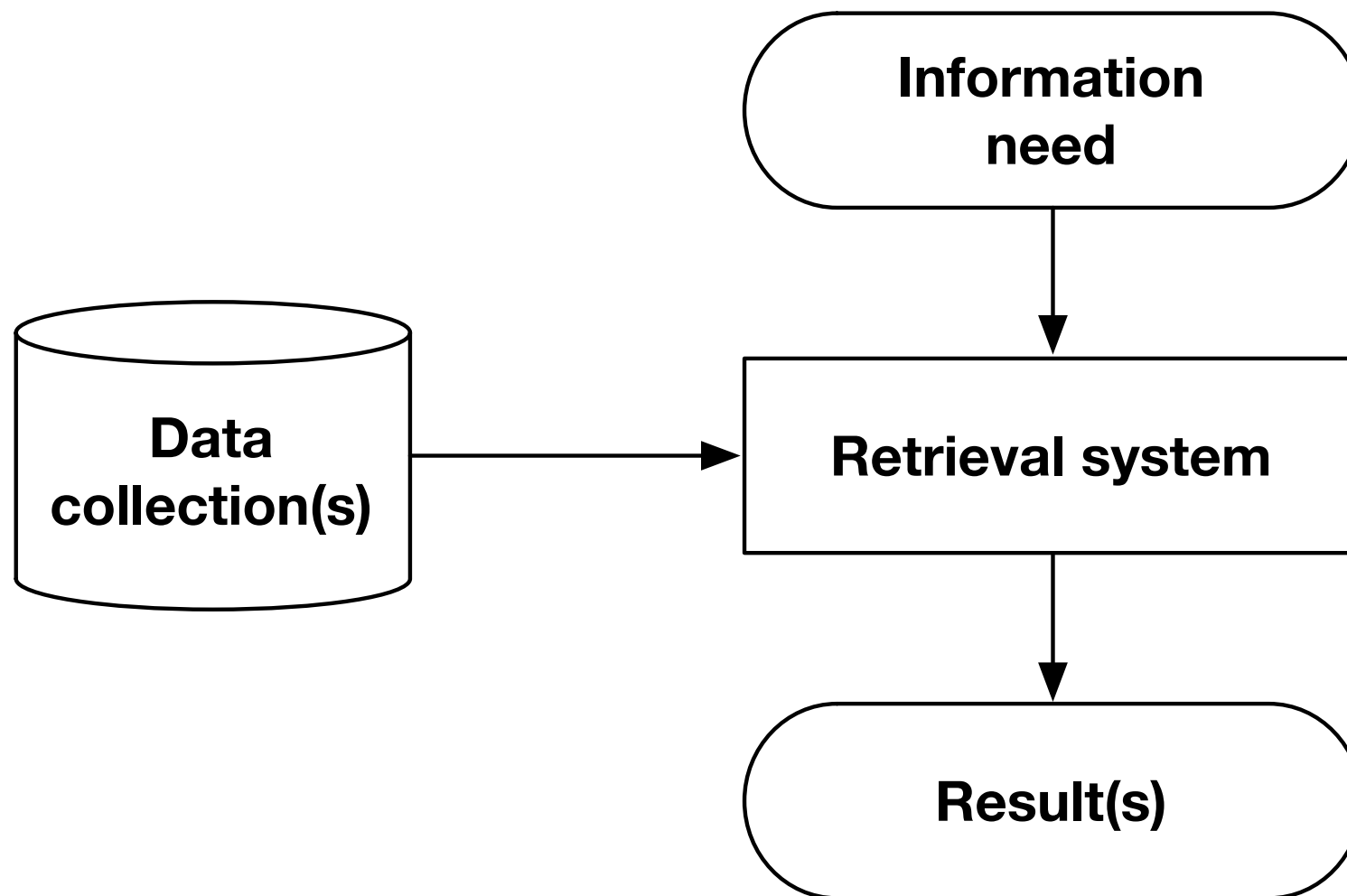
Distribution of web search queries [\[Lin et al. 2011\]](#)



Today's focus

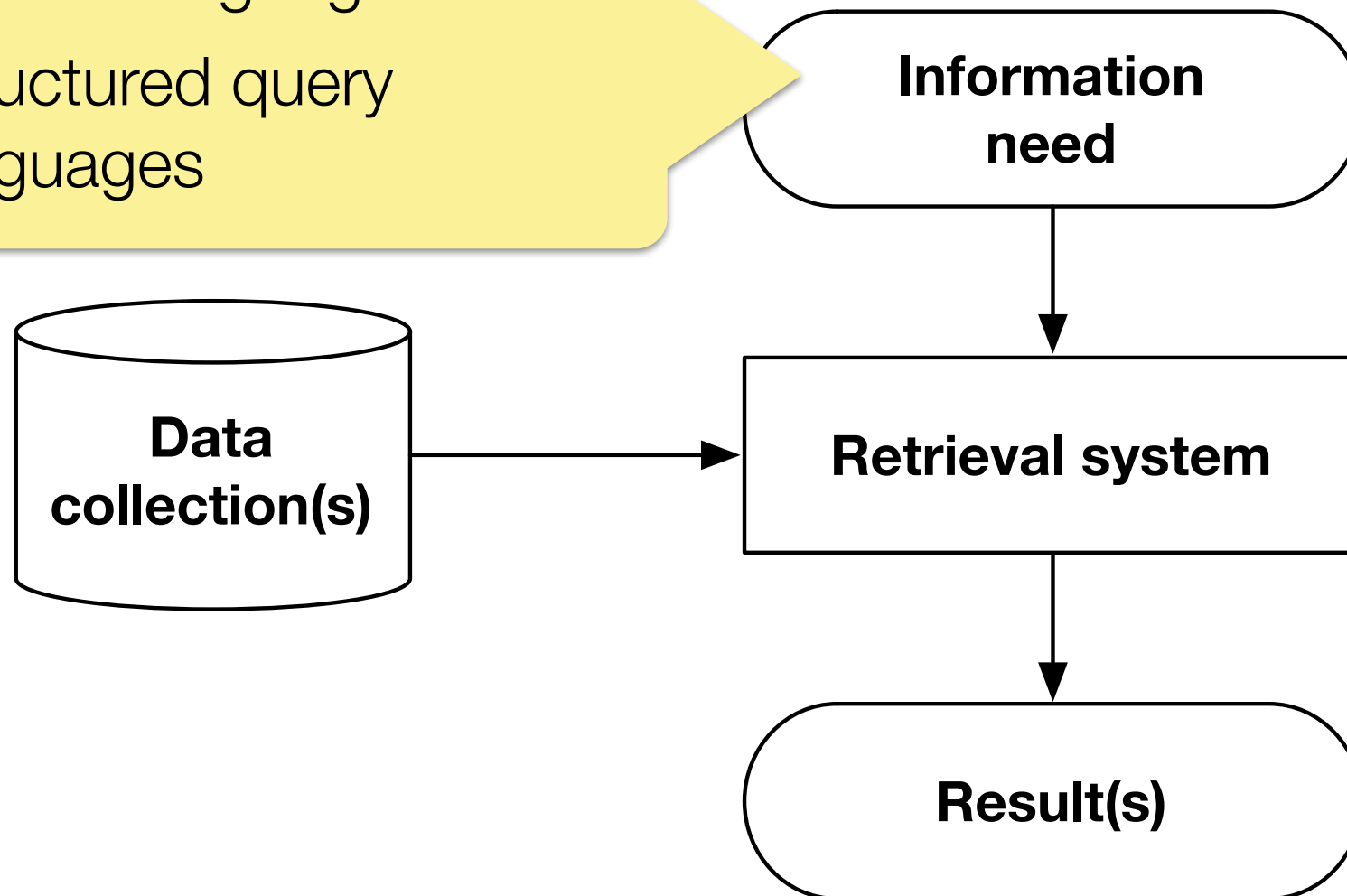
How to use KGs to improve information access.

Birds-eye view



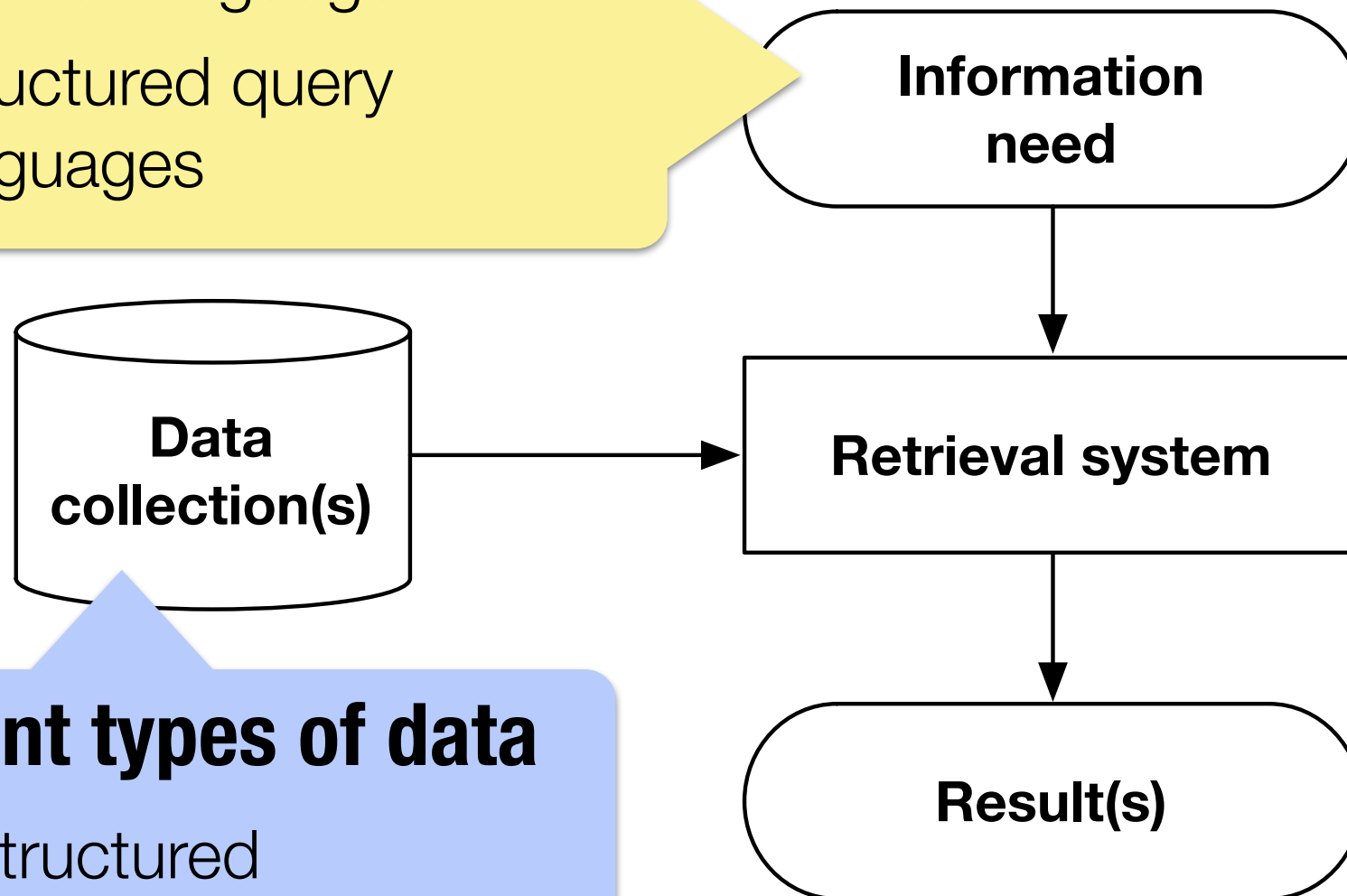
Many ways to express

- Keyword
- Keyword++
- Natural language
- Structured query languages



Many ways to express

- Keyword
- Keyword++
- Natural language
- Structured query languages

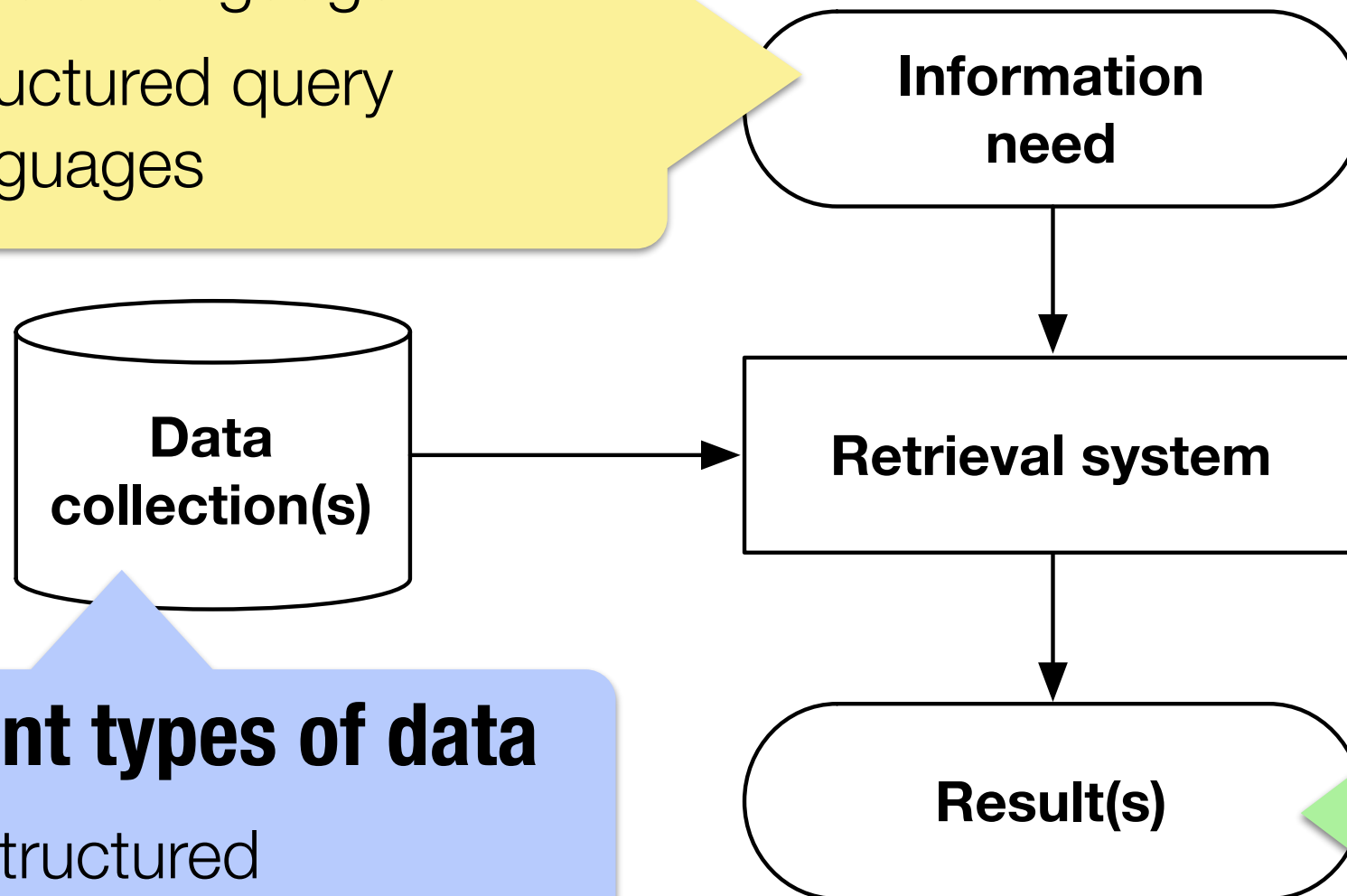


Different types of data

- Unstructured
- Semistructured
- Structured

Many ways to express

- Keyword
- Keyword++
- Natural language
- Structured query languages



Different types of data

- Unstructured
- Semistructured
- Structured

Result format

- Ranked list
- Tuples
- (Sub)graphs
- Natural language

Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO

Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



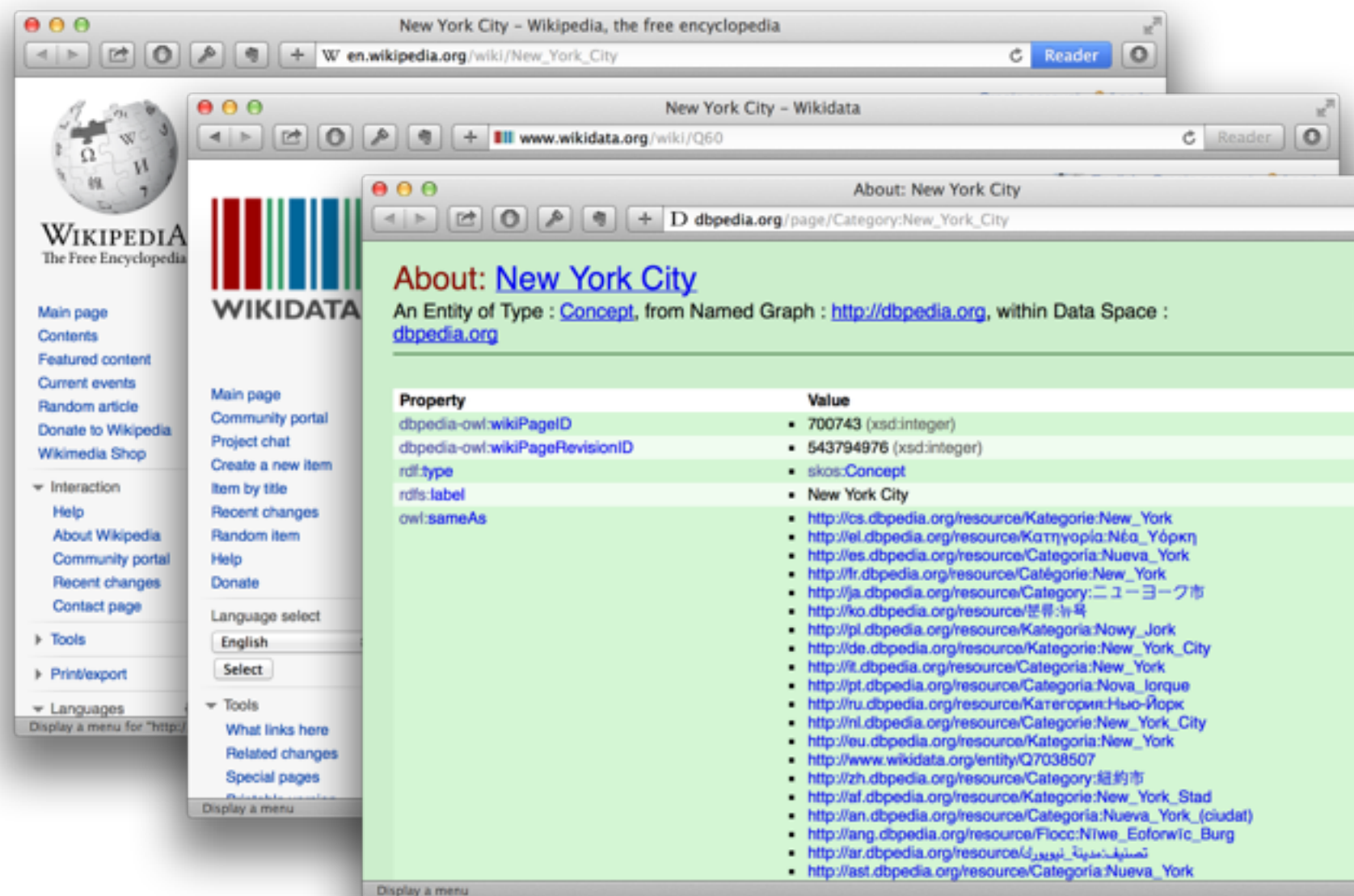
Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



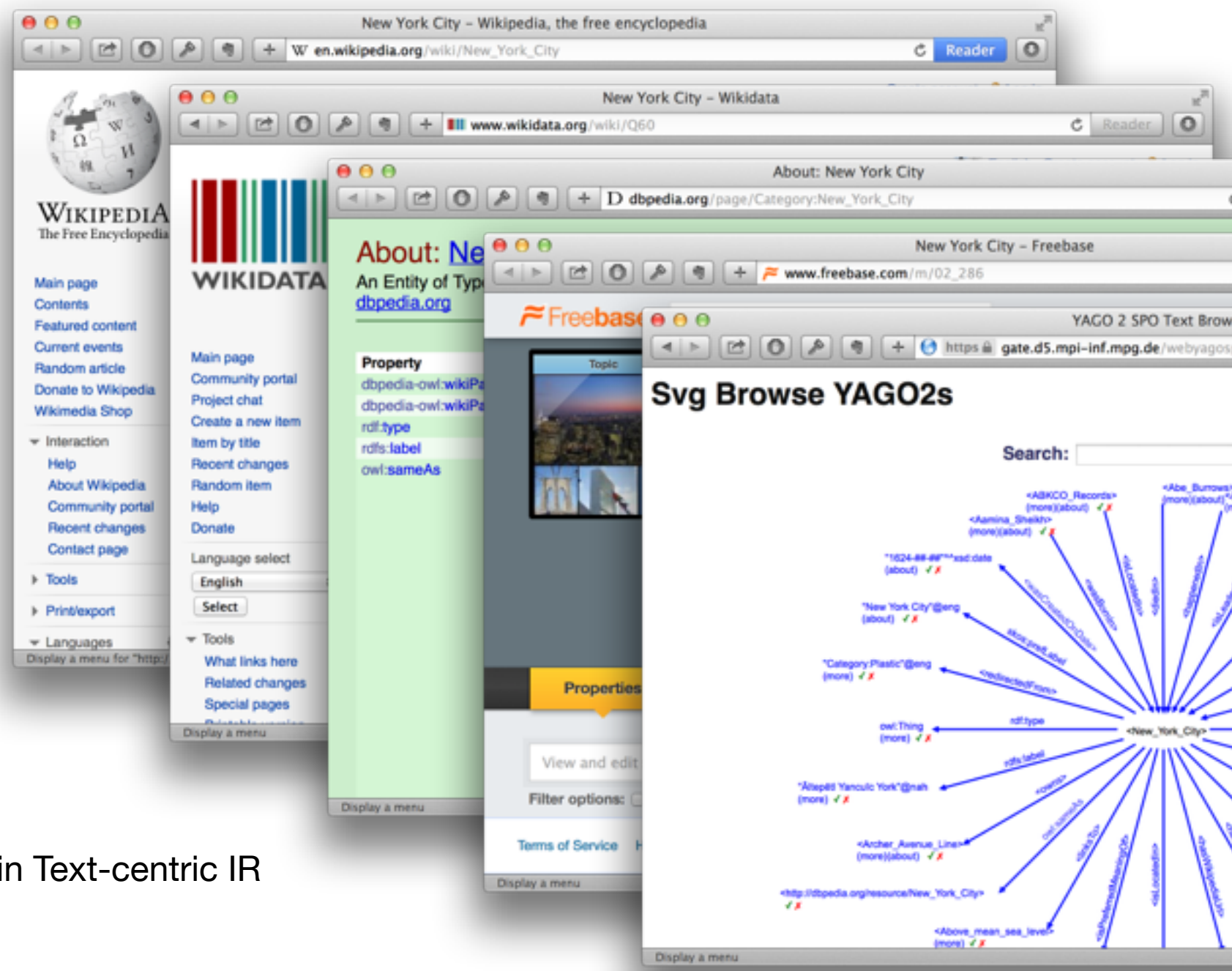
Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



DBpedia

- Extract structured information from Wikipedia
 - infoboxes, categories, and more
 - crowd-sourced community effort
- Open source
 - written in Scala, Java and VSP
 - Virtuoso Universal Server Operating system
- See <http://dbpedia.org/About>

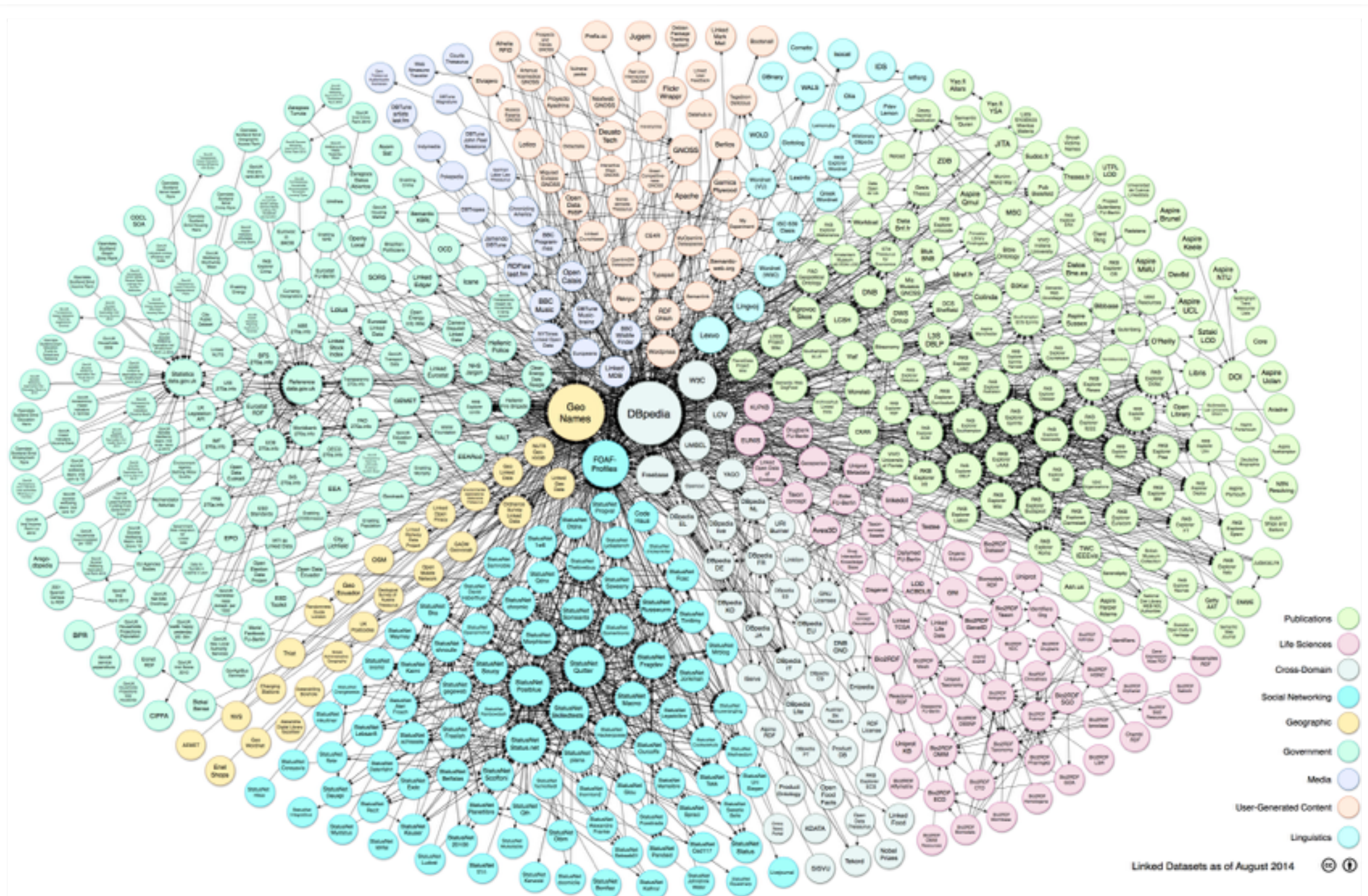
Freebase

- Initially seeded from high-quality open data
 - then maintained mainly by community
- Harvested from many sources
 - Wikipedia, MusicBrainz, and others.
- Acquired by Google in 2010 (GKG)
 - now in read-only mode
- See <http://www.freebase.com/>

YAGO

- Accuracy manually evaluated
 - confirmed accuracy of 95%
 - relations annotated with confidence values
- Anchored in Time and Space
 - Thematic domains (e.g. "music" or "science")
- Based on wikipedia, includes WordNet
- See <http://www.mpi-inf.mpg.de/yago-naga/yago/>

Linking Open Data (LOD)?



RDFa

- schema.org, sitemaps.org
 - used by Google, Bing, Yandex, Yahoo!, IPTC, etc.

RDFa

- schema.org, sitemaps.org
 - used by Google, Bing, Yandex, Yahoo!, IPTC, etc.

IMDb [The Wire \(TV Series 2002–2008\) - IMDb](#)
www.imdb.com/title/tt0306414/
★★★★★ Rating: 9.6/10 - 44,375 votes
Baltimore drug scene, seen through the eyes of drug dealers, and law enforcement.
Starring [Dominic West](#), [John Doman](#), [Deirdre Lovejoy](#).

ebay [Logitech Revue 097855070906 | eBay](#)
www.ebay.com/ctg/Logitech-Revue-/97019743
★★★★★ from 51 users - \$99.99 to \$190.00
eBay: The **Logitech Revue** is a media streamer with Google TV, which serves as a complete entertainment system. With its Ethernet Interface, this **Logitech**

ticketmaster [Washington Huskies Mens Basketball tickets, and ... - Ticketmaster](#)
www.ticketmaster.ca/Washington-Huskies-Mens-Basketball-tickets/art...
Results 1 - 10 of 19 – Find and buy **Washington Huskies Mens Basketball** ...
Sat 12 Nov **Washington Huskies Mens** ... - Alaska Airlines Arena at ...
Sun 13 Nov **Washington Huskies Mens** ... - Alaska Airlines Arena at ...
Mon 14 Nov **Washi**

iSaveurs! [Recette Tarte Tatin - testée et approuvée - calorie et ... - iSaveurs](#)
www.isaveurs.com/recette/recette_tart... - [Translate this page](#)
45 mins
il y a 1 jour – Recette de cuisine gourmande - Tarte Tatin - pour virginie63 :
J'adooooore j'adore, ...
http://www.isaveurs.com/recette/recette_tarte_tatin.php.

Menu

14:00 - 14:15	Introduction
14:15 - 14:45	Part 1 – Entity linking
14:45 - 15:30	Part 2 – Entity Representation and Retrieval
15:30 - 16:00	Coffee break
16:00 - 17:00	Part 3 – Utilizing KGs in Text-centric IR
17:00 - 17:30	Discussion and wrap-up