

Entity-Oriented Search

PRI 23/24 · Information Processing and Retrieval
M.EIC · Master in Informatics Engineering and Computation

Sérgio Nunes
Dept. Informatics Engineering
FEUP · U.Porto

Based on Entity-Oriented Search, Springer, K. Balog (2018)

Outline

- Overview of entity-oriented search
- Types of data sources
- Entity ranking
- Entity linking (briefly)

Entity-Oriented Search

Context

- Search is pervasive — on the web, on websites, on applications, on devices.
- Users expect to find a search box to help them find the information they need.
- Information retrieval deals with matching information needs with information objects.
- Queries are the expression of information needs.
- Queries may range from simple keywords [ps5 games] to full sentences ["Where can I buy a Playstation 5?"].
- The search engine responds with a ranked list of information objects.
- Traditionally these information objects were documents, but today's search engines return rich results pages beyond the classic "ten blue links" — *semantic search*.

Google Advanced Search Preferences Search Tips

Linux Google Search I'm Feeling Lucky

Searched the web for Linux. Results 1 - 10 of about 43,500,000. Search took 0.12 seconds.

LINUX - Most Reliable Linux Systems - Penguin Computing
www.penguincomputing.com Our High Quality LINUX Servers Power Start-ups and Fortune 500 Companies

Category: Computers > Software > Operating Systems > Linux

Welcome to internet.com's Linux/Open Source Channel
... Get Press Exposure! All Linux Devices | Enterprise Linux Today | Apache Today | PHPBuilder.com | BSD Today | Linuxnewbie | Linux Programming | Linux Today ...
www.internet.com/sections/linux.html - 42k - Cached - Similar pages

The Linux Home Page at Linux Online
... Linux is a free Unix-type operating system originally created by Linus Torvalds with the assistance of developers around the world. Developed under the GNU ...
Description: Comprehensive information and resources about the Linux Operating System.
Category: Computers > Software > Operating Systems > Linux
www.linux.org/ - 14k - Cached - Similar pages

Linux Today - Linux News On Internet Time.
... = Highlighted stories above -- Normal news below
-- Linux.com: Introduction To LDAP, ...
Description: It is the most frequently updated Linux News Site out there
Category: Computers > Software > Operating Systems > Linux > Linux in the Press
www.linuxtoday.com/ - 70k - Cached - Similar pages

Linux.com - A Means to World Liberation!
... Linux.com, Search Linux.com: ... site. JOLT LINUX.COM Journal of Linux Technology. The Latest of Linux.com, ...
Description: A great place to learn about Linux. A top quality Linux portal.
Category: Computers > Software > Operating Systems > Linux
www.linux.com/ - 38k - Cached - Similar pages

justlinux - The Complete Linux Guide
... Wed Oct 25, 2000 - 06:09 pm, ... More... More features.
PRESS RELEASES. LINUX NEWS. LinuxToday. ZDNet ...
Description: A complete guide to the rapidly growing Linux community. The website connects visitors to the Linux...
Category: Computers > Software > Operating Systems > Linux > Portal Sites
www.justlinux.com/ - 52k - Cached - Similar pages

Linux Central the /root for Linux resources
The root for Linux resources, including low cost CD's, official distributions, books, software, and applications.
Description: Sales of distributions, paraphernalia, apps, books and packages.
Category: Computers > Software > Operating Systems > Linux > Software > Retailers
www.linuxcentral.com/ - 53k - Cached - Similar pages

Enterprise Linux Today - Enterprise Linux Today
... David Brauer Freelance Writer for Enterprise Linux Today. Toyota to Save 3 Million A Year with the Help of Linux. ...
[www_eltoday.com/](http://eltoday.com/) - 32k - Cached - Similar pages

All Linux Devices - Linux News On Internet Time.
... on the Compaq iPAQ. The iPAQ is one of several handheld platforms upon which Linux developers are currently working to develop handheld Linux. Handhelds.org is ...
Description: Daily source of embedded Linux information. Linux Today format.
Category: Computers > Software > Operating Systems > Linux > Embedded
alllinuxdevices.com/ - 54k - Cached - Similar pages

LinuxPlanet - A Guide to the Linux Operating System

Google linux

All Images News Videos Maps Tools

About 3,450,000,000 results (0.76 seconds)

<https://www.linux.org> :
Linux.org
3 days ago — Linux.org · LFCS – Creating OpenLDAP Server on CentOS 7 · RPCS3 – Playstation 3 Emulator · Dolphin – Wii Emulator (even the newest versions of ...
Download Linux · Kali Linux · General Linux · Linux Hardware

People also ask :

- What is Linux and why it is used?
- Is Windows 10 better than Linux?
- Why Linux is so bad?
- Is Linux illegal?

<https://en.wikipedia.org/wiki/Linux> :
Linux - Wikipedia
A Linux-based system is a modular Unix-like operating system, deriving much of its basic design from principles established in Unix during the 1970s and 1980s.
OS family: Unix-like Initial release: September 17, 1991; 30 y...
Source model: Open source Default user interface: Unix shell (CLI); M...
List of Linux distributions · Linux distribution · Linux kernel · Fedora Linux

Top stories :

- Canaltech
Linux Mint 20.3 é lançado com novo visual, melhorias em apps e mais
18 hours ago
- Canaltech
Desenvolvedor pode resolver sozinho problema de 30 anos do Linux
2 days ago

More news

<https://ubuntu.com> :

Linux
Operating system

Linux is a family of open-source Unix-like operating systems based on the Linux kernel, an operating system kernel first released on September 17, 1991, by Linus Torvalds. Linux is typically packaged in a Linux distribution. [Wikipedia](#)

Initial release date: September 17, 1991
Programming languages: C, Assembly language
Original author: Linus Torvalds, Feral Interactive
Developer: Linus Torvalds, Feral Interactive
Update methods: KernelCare, dpkg, GNOME Software

Computer
Check disk space
Screenshots
Reboot

People also search for View 10+ more
mac Android Im NI

Google search for ["linux"] circa 2000 and today (2021).

Google Portugal

Cerca de 3 550 000 000 resultados (0,71 segundos)

[Portugal – Wikipédia, a encyclopédia livre](https://pt.wikipedia.org/wiki/Portugal)

Situado no extremo sudeste da Europa, Portugal Continental faz fronteira apenas com um outro país, Espanha a Este e a Norte, a Oeste e a Sul é limitado pelo ...

[Demografia de Portugal](#) · [Portugal Continental](#) · [Geografia de Portugal](#) · [Lisboa](#)

Notícias principais

- Público Covid-19. Portugal regista 39.074 novos casos e 25 mortes há 22 horas
- Observador Mapa Covid. Todas as regiões de Portugal a vermelho, em risco extremo há 14 horas
- O Jornal Económico "Em Portugal não há uma estratégia coordenada com a internacionalização" há 40 minutos
- Diário de Notícias Porque investem os estrangeiros em Portugal? há 12 horas

[Mais notícias](#)

<https://www.portugal.gov.pt>

República Portuguesa: XXII Governo
Página Oficial do Governo de Portugal - República Portuguesa.

Vídeos

- Portugal prolonga teletrabalho mas alivia condições de ... YouTube · euronews (em português) há 18 horas
- Legislativas 22 - Debates TVI/CNN Portugal Episódio 2 - de ... RTP há 2 dias
- «Quando fui embora de Portugal, dei xeito para trás a Jaciara ... TVI há 1 dia

[Ver tudo](#)

Google tripas à moda do porto

Cerca de 203 000 resultados (0,50 segundos)

Receitas

Prato	Preparador	Classificação	Duração
Tripas à moda do porto	Petitchef	4,4 ⭐⭐⭐⭐ (30)	2 h
Tripas à moda do porto receta - Aprende a...	Solteiros contra Casados	Sem classificações	4 h
Tripas à moda do Porto	Pingo Doce	2,0 ⭐⭐⭐⭐ (21)	1 h 30 min

[Mostrar mais](#)

<https://pt.petitchef.com/.../receitas-com-cenoura>

Tripas à moda do porto - Receita Petitchef
Num tacho grande pique uma cebola grande e um dente alho, junte azeite e leve a refogar, junte a carne de vaca partida aos bocados e um pouco de vinho, deixe ...
★★★★★ Classificação: 4,4 · 30 votos · 2 h

Vídeos

- Tripas à Moda do Porto do Chef Hélio Loureiro | Receitas com ... YouTube · Continente 30/03/2021
- Tripas à moda do porto YouTube · Ana Cozinha 28/08/2016
- Tripas à Moda do Porto | Praça da Alegria | RTP YouTube · RTP Receitas 21/11/2018

[Ver tudo](#)

<https://solteiroscontracasados.com/pagina-de-artigos>

Tripas à moda do porto receita - Aprende a fazer tripas
20/08/2020 — Ingredientes das tripas à moda do porto · 1kg de tripas (também conhecido como folhos); · 1kg de feijão manteiga demolido; · 1 mão de vitela; · 1

Rich search engine results pages (SERPs), containing links to documents, information boxes, direct display of entities, facts, and other structured results.

Google fcporto results

Cerca de 7 410 000 resultados (0,51 segundos)

Futebol Clube do Porto

1º em Primeira Liga

JOGOS	NOTÍCIAS	CLASSIFICAÇÕES	JOGADORES
Porto vs Benfica (3-1) - TER. Quinta, 30/12		Taça de Portugal - Oitavos de final (Porto vs Benfica - 3-0) - TER. 23/12	
Vizela vs Porto (0-4) - TER. 19/12		Taça da Liga - Fase de grupos - Dia de jogo 3 de 3 (Porto vs Rio Ave - 1-0) - TER. 15/12	
Porto vs Braga (1-0) - TER. 12/12		Liga dos Campeões - Fase de grupos - Dia de jogo 6 de 6 (Porto vs Atlético Madrid - 1-3) - TER. 07/12	

[Ver mais](#)

<https://www.timeout.com> > porto - Traduzir esta página

The 30 best things to do in Porto right now - Time Out

07/10/2021 — What is it? It is one of the biggest (if not the biggest) ex-libris of the city. And the tour through Invicta (another name for Porto) ...

<https://www.thecommonwanderer.com> > ... - Traduzir esta página

16 Unmissable Things to do in Porto, Portugal [2021] - The ...

15/02/2020 — 16 AWESOME THINGS TO DO IN PORTO - THE PORTO CARD | A PORTO ESSENTIAL · #1 TAKE A SIGHTSEEING CRUISE DOWN THE DOURO RIVER · #2 TAK...

[Feedback](#)

Google porto things to see

Cerca de 48 000 000 resultados (1,19 segundos)

Principais atrações em Porto

- Livraria Lello** (4.2 ★★★★★) (40.669) Livraria ornada com escadaria vermelha
- São Bento** (4.7 ★★★★★) (2.449) Estação do séc. XIX com azulejos ornados
- Ponte Luís I** (4.8 ★★★★★) (56.504) Ponte icónica em metal sobre o Douro

[Mais coisas a fazer](#)

<https://www.timeout.com> > porto - Traduzir esta página

The 30 best things to do in Porto right now - Time Out

07/10/2021 — What is it? It is one of the biggest (if not the biggest) ex-libris of the city. And the tour through Invicta (another name for Porto) ...

<https://www.thecommonwanderer.com> > ... - Traduzir esta página

16 Unmissable Things to do in Porto, Portugal [2021] - The ...

15/02/2020 — 16 AWESOME THINGS TO DO IN PORTO - THE PORTO CARD | A PORTO ESSENTIAL · #1 TAKE A SIGHTSEEING CRUISE DOWN THE DOURO RIVER · #2 TAK...

[Feedback](#)

As pessoas também perguntam

What should I not miss Porto?

What should I see in Porto?

Is 2 days enough for Porto?

What is Porto very famous for?

[Feedback](#)

<https://www.thecrazytourist.com> > ... - Traduzir esta página

25 Best Things to Do in Porto (Portugal) - The Crazy Tourist

25 Best Things to Do in Porto (Portugal) ; Cais da Ribeira ; Cais da Ribeira ; Serralves

[Display a menu](#)

Vizinhanças: Ribeira, Bela Vista, Bolhão, Pasteleira, Prelada, Antas, MAIS

Faculdades e universidades [Ver mais de 5](#)

POR **U-POR** **IP-PORTO** **U-PORTUGAL**

More examples of "semantic search".

Knowledge Bases and Entities

- A primary component enabling these advanced search services is the availability of large-scale structured knowledge repositories, called knowledge bases.
- Knowledge bases organize information around specify things or objects, called "entities".
- An entity is a uniquely identifiable object or thing, characterized by its name(s), type(s), attributes, and relationships to other entities.
- Common types of entities include: people, organizations, products, locations, and events.

Properties of Entities

- **Unique identifier:** entities need to be uniquely identifiable. There must be a one-to-one correspondence between each entity identifier and the object it represents. Examples include: username, email address, URI, Wikipedia page ID, etc.
- **Name(s):** entities are known and referred to by their name. Unlike IDs, names do not uniquely identify entities; multiple entities may share the same name; and the same entity may be known by more than a single name.
- **Type(s):** entities may be categorized into multiple entities types. Types work as semantic categories that group together entities with similar properties. The set of possible entity type is often organized in a hierarchical structure, e.g. "scientist" is a subtype of "person".
- **Attributes:** the characteristics or features of an entity are described by a set of attributes. Different entities have different sets of attributes. Some of the characteristics may be entities themselves, in this we cases they are not treated as attributes but as relationships, e.g. "place of birth" links to a location entity. Attributes always have literal values and optionally may also include data type information.
- **Relationships:** describe how two entities are associated to each other. Relationships may also be seen as "typed links" between entities.

Representing Properties of Entities

- Information about entities can be represented and stored in semi-structured or in structured form.
 - Wikipedia is an example of a knowledge repository that organizes information about entities and their attributes and relationships in a semi-structured form.
 - To adopt a structured form, a knowledge representation model is necessary. The Resource Description Framework (RDF) is a standard way to describe entities, where an entity is represented as a set of RDF statements.
- A knowledge base (KB) is a structured knowledge repository that contains a set of facts (assertions) about entities.
- Entities in a knowledge base may be seen as nodes in a graph, with relationships between them as edges. Thus, knowledge bases are also referred to as knowledge graphs.

About: University of Porto

An Entity of Type: [Public university](#), from Named Graph: <http://dbpedia.org>, within Data Space: [dbpedia.org](#)

The University of Porto (Universidade do Porto) is a Portuguese public university located in Porto, and founded on 22 March 1911. It is the second largest Portuguese university by number of enrolled students, after the University of Lisbon, and has one of the most noted research outputs in Portugal.



Property	Value
dbo:abstract	<ul style="list-style-type: none">The University of Porto (Universidade do Porto) is a Portuguese public university located in Porto, and founded on 22 March 1911. It is the second largest Portuguese university by number of enrolled students, after the University of Lisbon, and has one of the most noted research outputs in Portugal. (en)
dbo:affiliation	<ul style="list-style-type: none">dbr:CESAER
dbo:city	<ul style="list-style-type: none">dbr:Porto
dbo:country	<ul style="list-style-type: none">dbr:Portugal
dbo:motto	<ul style="list-style-type: none">Virtus Unita Fortius Agit
dbo:numberOfPostgraduateStudents	<ul style="list-style-type: none">8235 (xsd:nonNegativeInteger)
dbo:numberOfStudents	<ul style="list-style-type: none">30640 (xsd:nonNegativeInteger)
dbo:numberOfUndergraduateStudents	<ul style="list-style-type: none">22405 (xsd:nonNegativeInteger)
dbo:officialSchoolColour	<ul style="list-style-type: none">Gold
dbo:other	<ul style="list-style-type: none">742 (xsd:integer)
dbo:thumbnail	<ul style="list-style-type: none">wiki-commons:Special:FilePath/Fonte_dos_leões_e_reitoria.jpg?width=300

Example of the "University of Porto" entity in DBpedia
https://dbpedia.org/page/University_of_Porto

```
<http://dbpedia.org/resource/University_of_Porto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/University> .  
<http://dbpedia.org/resource/University_of_Porto> <http://xmlns.com/foaf/0.1/homepage> <https://sigarra.up.pt/up/en/WEB_PAGE.INICIAL%7C2=www.up.pt> .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/ontology/country> <http://dbpedia.org/resource/Portugal> .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/ontology/city> <http://dbpedia.org/resource/Porto> .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/property/wikiPageUsesTemplate> <http://dbpedia.org/resource/Template:Color_box> .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/ontology/wikiPageWikiLink> <http://dbpedia.org/resource/Higher_education_in_Portugal> .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/property/rector> "António Sousa Pereira"@en .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/ontology/motto> "Virtus Unita Fortius Agit" .  
<http://dbpedia.org/resource/University_of_Porto> <http://xmlns.com/foaf/0.1/name> "Universidade do Porto"@en .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/ontology/abstract> "The University of Porto (Universidade do Porto) is a Portuguese public  
university located in Porto, and founded on 22 March 1911. It is the second largest Portuguese university by number of enrolled students, after the University of  
Lisbon, and has one of the most noted research outputs in Portugal."@en .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/property/country> <http://dbpedia.org/resource/Portugal> .  
<http://dbpedia.org/resource/University_of_Porto> <http://dbpedia.org/property/nativeName> "Universidade do Porto"@en .
```

Excerpt from the RDF definition of the "University of Porto" entity in DBpedia in N-Triples
https://dbpedia.org/page/University_of_Porto

Semantic Web

- The Semantic Web is a term coined by Tim Berners-Lee in 2001 referring to an envisioned extension to the original web.
- While the original web is a medium of documents for people — a Web of Documents.
- The semantic web is meant to be a medium for machines (intelligent agents) that have access to data and rules for reasoning about the data — a Web of Data.
- Semantic Web technologies include: URI, XML, RDF, various serializations of RDF (Turtle, N-Triples, RDFa, etc), SPARQL, OWL, etc — commonly referred to as the Semantic Web Stack.
- This vision is yet (?) to come true, however it has contributed to the development and publication of structured data in an unprecedented scale.

Entity-Oriented Search

- Entity-Oriented Search is the search paradigm of organizing and accessing information centered around entities, and their attributed and relationships.
- From a user perspective, entities are natural units for organizing information. Allowing users to interact with specific entities offers a richer and more effective user experience than what is provided by conventional document-based retrieval systems.
- From a machine perspective, entities allow for a better understanding of search queries, of document content, and even of users. Entities enable search engines to be more effective.

Entity-Oriented Search

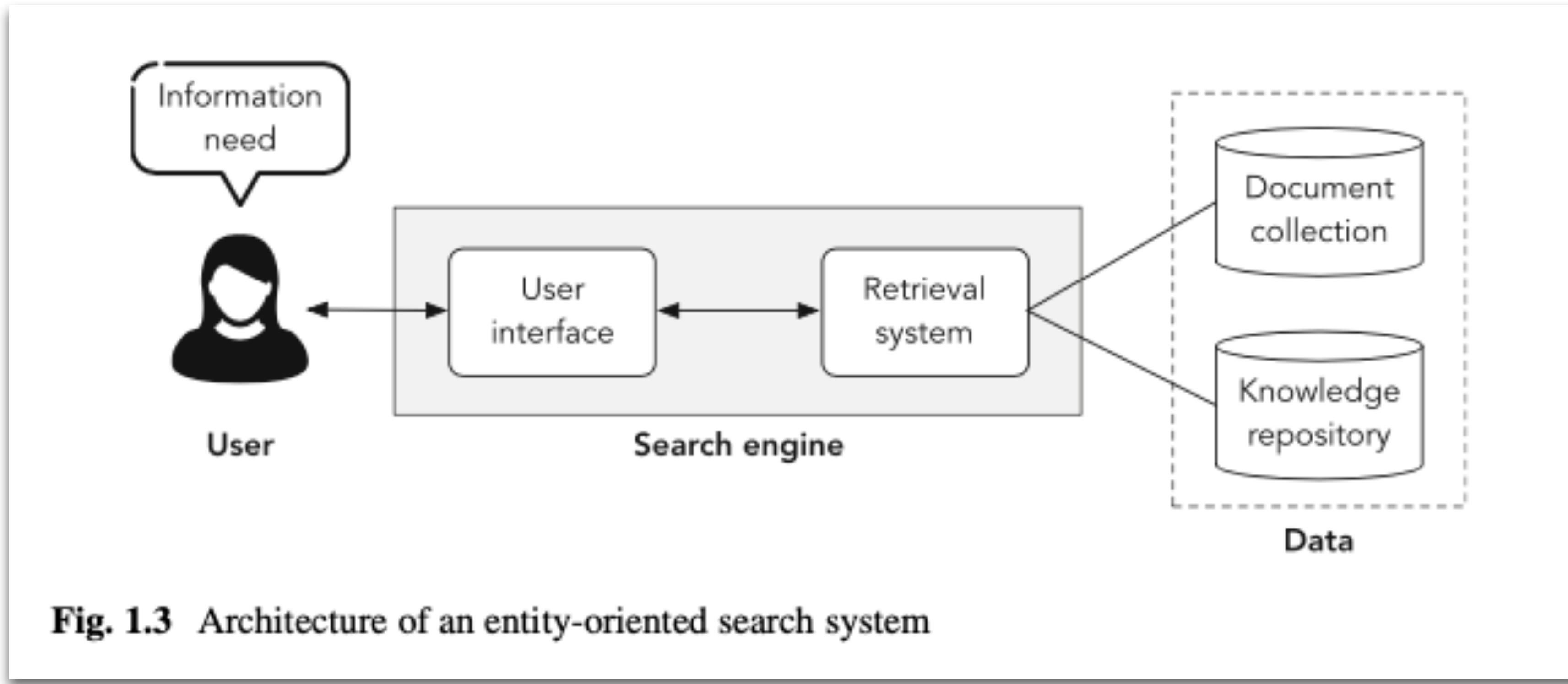


Fig. 1.3 Architecture of an entity-oriented search system

Users and Information Needs

- Users articulate their information needs in many different ways.
- **Keyword queries**, also known as "free text queries", have become the most common approach with the advent of the web. Keyword queries are easy to formulate but imprecise.
- **Structured queries**, structured data sources are traditionally queried using formal query languages (e.g., SQL, SPARQL). These queries are very precise and usually intended for expert users with well-defined and precise information needs.
- **Keyword++ queries**, correspond to keyword queries that are complemented with additional structural elements, e.g. [retrieval site:fe.up.pt].
- **Natural language queries**, are formulated using natural language, the same way one would express it in a conversation, e.g. ["What is the capital of Nepal?"].
- **Zero-queries**, the traditional information access is reactive — the search system responds to a user-initiated query. Proactive systems, anticipate the user's information need, without requiring the user to issue a query. In practice, the context of the user is used to infer information needs.

Data

- Data can be grouped in three types.
- **Unstructured data**, which can be found in vast quantities in a variety of forms: web pages, spreadsheets, emails, tweets, etc. All these may be treated as textual documents, i.e. a sequence of words.
- **Semi-structured data**, is characterized by the lack of rigid, formal structure. Typically, it contains tags or other markup to separate textual content from semantic elements, e.g. HTML data.
- **Structured data**, adheres to a predefined (fixed) schema and is typically organized in a tabular format (e.g., relational databases). The schema defines how the data is organized and imposes constraints to ensure consistency.

Tasks in Entity-Oriented Search

- **Entity Retrieval – entities as the unit of retrieval**, between 40% to 70% of the queries in web search mention or target specific entities. Such queries are better answered by returning a ranked list of entities, as opposed to a list of documents. Challenges involved:
 - 1) how to represent information needs;
 - 2) how to represent entities;
 - 3) how to match those representations.
- **Entity Linking – entities for knowledge representation**, recognizing mentions of entities in text and associating these mentions with the corresponding entities in knowledge bases.
- **Entities for an enhanced user experience**, besides providing meaningful retrieval as information organization units, entities can improve the user experience throughout the entire search process.

Data Sources

Data Types and Search

→ Recall the basic data types of data sources.



Table 2.1 Comparison of unstructured, semi-structured, and structured data search

	Unstructured	Semi-structured	Structured
Unit of retrieval	Documents	Objects	Tuples
Schema	No	Self-describing	Fixed
Queries	Keyword	Keyword++	Formal languages

Knowledge Bases

- A knowledge base is comprised of a large set of assertions about the world.
- When the emphasis is on the relationships between the entities, the term knowledge graph is commonly used.
- Resource Description Framework (RDF) is a language designed to describe "things", which are referred to as resources.
- Each resource is assigned a Uniform Resource Identifier (URI), making it uniquely and globally identifiable.
- Each RDF statement is a triple, consisting of subject, predicate, and object components
 - Subject, always a URI, denoting a resource;
 - Predicate, always a URI, corresponding to a relationship or property of the subject resource;
 - Object, either a URI (referring to another resource) or a literal.

Knowledge Bases

Michael Schumacher (born 3 January 1969) is a retired German racing driver, who raced in Formula One for Ferrari.

<dbr:Michael_Schumacher>	<foaf:name>	"Schumacher, Michael"
<dbr:Michael_Schumacher>	<dbo:birthPlace>	<dbr:West_Germany>
<dbr:Michael_Schumacher>	<dbo:birthDate>	"1969-01-03"
<dbr:Michael_Schumacher>	<rdf:type>	<dbo:RacingDriver>
<dbr:Michael_Schumacher>	<dct:subject>	<dbc:Ferrari-Formula-One-drivers>

Knowledge Bases

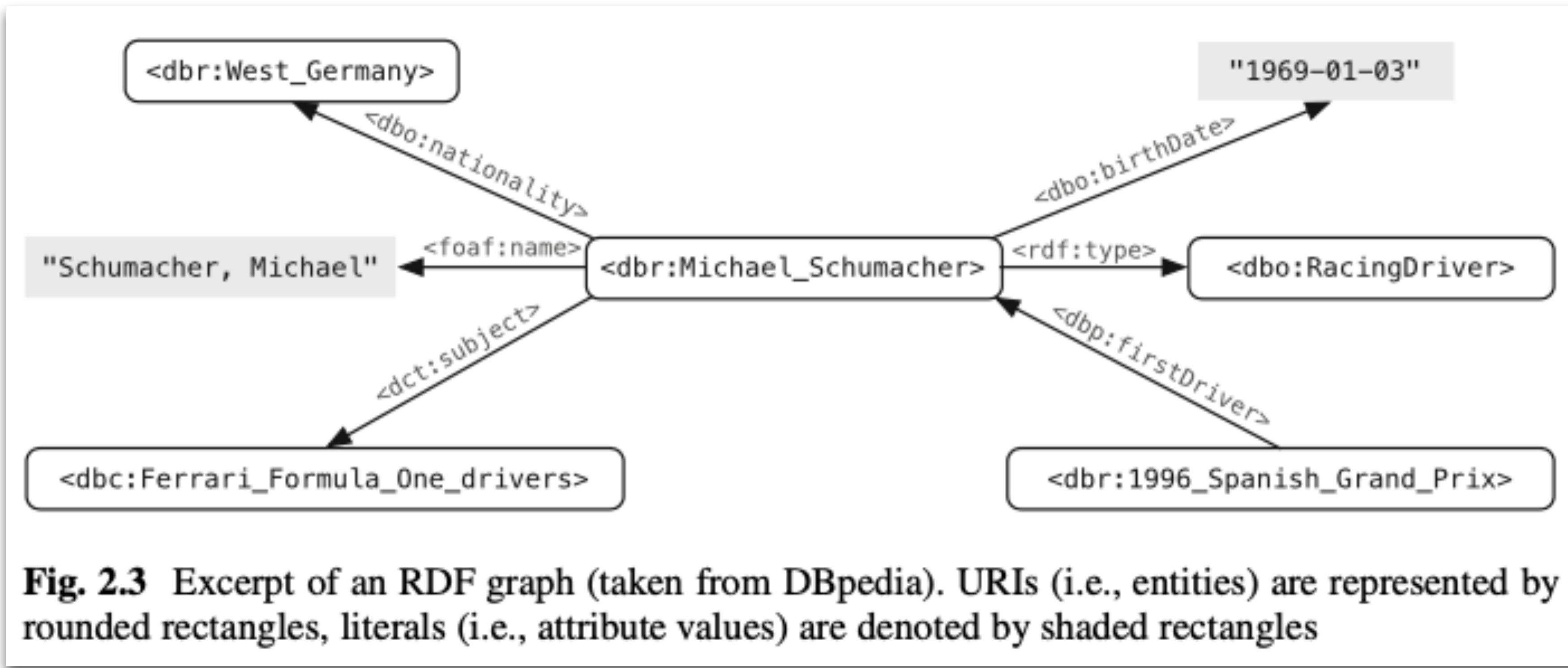


Fig. 2.3 Excerpt of an RDF graph (taken from DBpedia). URIs (i.e., entities) are represented by rounded rectangles, literals (i.e., attribute values) are denoted by shaded rectangles

Related Technologies

- RDF describes the instance level in the knowledge base.
- RDFS and OWL are vocabularies for ontological modeling
 - An ontology is a means to formalize knowledge. Building blocks of an ontology include classes, instances, relations, attributes, restrictions, and rules and axioms.
- Serializations for RDF data: Notation-3, Turtle, N-Triples, RDFa, and RDF/JSON.
- SPARQL is structured query language for retrieving and manipulating RDF data.
- Triplestores are special-purpose databases designed for storing and querying RDF data.

Public Knowledge Bases

- DBpedia
 - Extracted from Wikipedia (mostly from infoboxes) using a set of manually constructed mapping rules.
 - Community effort, users collaboratively create and edit the mapping rules.
 - Available in multiple languages.
 - Contains over 5 million entities (English).

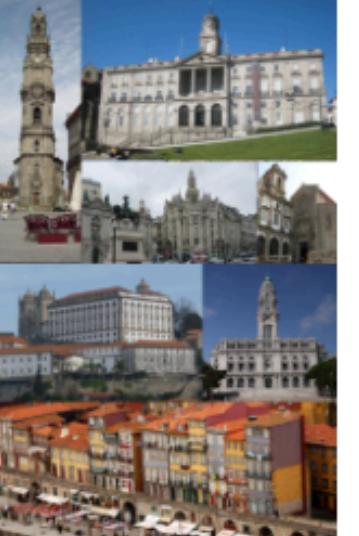
DBpedia

DBpedia [Browse using](#) [Formats](#) [Faceted Browser](#) [Sparql Endpoint](#)

About: Porto

An Entity of Type: [Municipality](#), from Named Graph: <http://dbpedia.org>, within Data Space: [dbpedia.org](#)

Porto or Oporto (Portuguese pronunciation: [ˈportu]) is the second-largest city in Portugal and one of the Iberian Peninsula's major urban areas. Porto city proper, which is the entire municipality of Porto, is small compared to its metropolitan area, with an estimated population of just 215,945 people in a municipality with only 41.42 km². Porto's metropolitan area has around 1.7 million people (2019) in an area of 2,395 km² (925 sq mi), making it the second-largest urban area in Portugal. It is recognized as a global city with a Gamma + rating from the Globalization and World Cities Research Network.



Property	Value
dbo:PopulatedPlace/areaTotal	• 41.42
dbo:abstract	• Porto or Oporto (Portuguese pronunciation: [ˈportu]) is the second-largest city in Portugal and one of the Iberian Peninsula's major urban areas. Porto city proper, which is the entire municipality of Porto, is small compared to its metropolitan area, with an estimated population of just 215,945 people in a municipality with only 41.42 km ² . Porto's metropolitan area has around 1.7 million people (2019) in an area of 2,395 km ² (925 sq mi), making it the second-largest urban area in Portugal. It is recognized as a global city with a Gamma + rating from the Globalization and World Cities Research Network. Located along the Douro River estuary in northern Portugal, Porto is one of the oldest European centres, and its core was proclaimed a World Heritage Site by UNESCO in 1996, as "Historic Centre of Porto, Luiz I Bridge and Monastery of Serra do Pilar". The historic area is also a National Monument of Portugal. The western part of its urban area extends to the coastline of the Atlantic Ocean. Its settlement dates back many centuries, when it was an outpost of the Roman Empire. Its combined Celtic-Latin name, Portus Cale, has been referred to as the origin of the name Portugal, based on transliteration and oral evolution from Latin. In Portuguese, the name of the city includes a definite article: o Porto ("the port" or "the harbor"), which is where its English name "Oporto" comes from. Port wine, one of Portugal's most famous exports, is named after Porto, since the metropolitan area, and in particular the cellars of Vila Nova de Gaia, were responsible for the packaging, transport, and export of fortified wine. In 2014 and 2017, Porto was elected The Best European Destination by the Best European Destinations Agency. Porto is on the Portuguese Way path of the Camino de Santiago. (en)

Public Knowledge Bases

- Wikidata
 - Operated by the Wikimedia Foundation.
 - Its goal is to provide the same information as Wikipedia, but in a structured format.
 - Wikidata considers "claims" not "facts"
 - Each claim must be supported by a reference;
 - Claims can contradict each other and coexist, thereby allowing opposing views to be expressed (e.g., different political positions).

Wikidata

Wikidata English Not logged in Talk Contributions Create account Log in

Item Discussion Read View history Search Wikidata

Porto (Q322792)

District of Portugal
Porto District

In more languages Configure

Language	Label	Description	Also known as
English	Porto	District of Portugal	Porto District
Portuguese	Porto	distrito de Portugal	distrito do Porto distrito portuense Distrito do Porto
French	Porto	district du Portugal	
Spanish	Distrito de Oporto	distrito portugués	Distrito de Porto Distrito do Porto Oporto

All entered languages

Statements

instance of district of Portugal
native label Distrito do Porto (Portuguese)
country Portugal

Proprietary Knowledge Bases

- Google Knowledge Graph
- Facebook Entity Graph
- Microsoft Satori

Entity Ranking

Ad Hoc Entity Retrieval Task

- Ad hoc entity retrieval is the task of answering queries with a ranked list of entities.
- "Ad hoc" refers to the standard form of retrieval in which the user, motivated by an ad hoc (extemporary) information need, initiated the search process by formulating and issuing a query.

Table 3.1 Example entity search queries taken from various benchmarking campaigns [4]

Query
martin luther king
disney orlando
Apollo astronauts who walked on the Moon
Winners of the ACM Athena award
EU countries
Hybrid cars sold in Europe
birds cannot fly
Who developed Skype?
Which films starring Clint Eastwood did he direct himself?

Main Strategy

- Model the "entity retrieval" problem as a "document retrieval" problem and take advantage of all the existing body of work.
- Create an entity description, or "profile document", for each entity in the catalog, containing all the existing "knowledge" about that entity, based on the available data.
- Once these entity descriptions are created, they can be indexed and ranked using existing document retrieval algorithms.
- Involves two steps, discussed in the next slides:
 - Constructing the profile documents;
 - Ranking the profile documents.

Constructing Entity Descriptions

- First step is to create a "profile document" with all the information we have about that entity.
- This will serve as the textual representation of the given entity, the "entity description".
- Entity descriptions can be assembled by considering the textual content, from a document collection, in which the entities occur.
- In some cases, such descriptions may already be easily available – e.g., Wikipedia page of an entity. Also, there is a lot of information about entities organized and stored in knowledge bases.
- Three possible data sources: unstructured, semi-structured and structured.
- The goal is to estimate a term count associated with each entity.

Entity Components

- The estimative of the number of times a term is associated with an entity can be used to compute standard components, in an analogous way to the document model.
- **Entity length**, is the total number of terms in the entity description.
- **Term frequency (TF)**, is the normalized term count (by length) in the entity description.
- **Entity frequency (EF)**, is the number of entities in which the term occurs.
- **Inverse entity frequency (IEF)**, is the log normalized ratio between the total number of entities in the catalog, and the entity frequency.

Entity Representations from Unstructured Data

- This is the scenario where we want to find entities in arbitrary document collections.
- An approach is to have documents annotated with the entities that are references in them, and use these documents as a proxy to connect terms and entities.
- These annotations may be provided by human editors or be automated (e.g., entity linking).
- In this scenario, the estimated entity term count is obtained by considering each document, and adding the number of co-occurrences between a term and an entity, weighted by the strength of the association between the entity and the document.

Entity Representations from Semi-Structured Data

- A significant portion of the web content is already organized around entities, e.g., Wikipedia page about a person, IMDb page about a movie or an actor.
- The standard way of incorporating the internal document structure into the retrieval model is through the use of document fields, where document segments correspond to fields — e.g., title, summary, introduction, etc.
- The standard statistics are now computed at field-level, e.g. the number of times a term appears in a field of the description of an entity.
- A special "catch-all" field is introduced to
 - (1) quickly filter entities in a first-pass retrieval;
 - (2) since entity description fields are often sparse, improve ranking by combining field-level scores with entity-level (the catch-all field) scores.

Entity Representations from Semi-Structured Data



The screenshot shows the movie page for "The Matrix" (1999) on IMDb. At the top, there are links for "FULL CAST AND CREW", "TRIVIA", "USER REVIEWS", "IMDbPro", and "MORE". Below that is a "SHARE" button. The main title "The Matrix (1999)" is displayed with a yellow star rating of "8.7 / 10" and a count of "1,235,392". A "Rate This" button is also present. The movie's rating is R, it has a runtime of 2h 16min, and it is an Action, Sci-Fi film released on 31 March 1999 (USA). The plot summary states: "A computer hacker learns from mysterious rebels about the true nature of his reality and his role in the war against its controllers." Below the plot is a list of credits: Directors: Lana Wachowski (as The Wachowski Brothers), Lilly Wachowski (as The Wachowski Brothers); Writers: Lilly Wachowski (as The Wachowski Brothers), Lana Wachowski (as The Wachowski Brothers); Stars: Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss. There is a link to "See full cast & crew". At the bottom, there are sections for Metascore (73 from metacritic.com), Reviews (3,655 user | 312 critic), and Popularity (286). A yellow banner at the bottom highlights "Top Rated Movies #18" and "Won 4 Oscars".

Fig. 3.4 Web page of the movie THE MATRIX from IMDb (<http://www.imdb.com/title/tt0133093/>)

Table 3.3 Fielded entity description created for THE MATRIX, corresponding to Fig. 3.4

Field	Content
Name	The Matrix
Genre	Action, Sci-Fi
Synopsis	A computer hacker learns from mysterious rebels about the true nature of his reality and his role in the war against its controllers
Directors	Lana Wachowski (as The Wachowski Brothers), Lilly Wachowski (as The Wachowski Brothers)
Writers	Lilly Wachowski (as The Wachowski Brothers), Lana Wachowski (as The Wachowski Brothers)
Stars	Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss
Catch-all	The Matrix Action, Sci-Fi A computer hacker learns from mysterious rebels about the true nature of his reality and his role in the war against its controllers. Lana Wachowski (as The Wachowski Brothers), Lilly Wachowski (as The Wachowski Brothers) Lilly Wachowski (as The Wachowski Brothers), Lana Wachowski (as The Wachowski Brothers) Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss

Entity Representations from Structured Data

- There is a growing number of large general-purpose publicly available knowledge bases.
- In knowledge bases information is organized using the RDF data model, specifically
 - each entity is uniquely identified by its URI (Uniform Resource Identifier);
 - its properties are described in the form of subject-predicate-object (SPO) triples.
- To represent an entity from a knowledge base, we consider the immediate vicinity of the entity node – i.e. the SPO triples where the entity appears either as subject or object.
 - <<http://dbpedia.org/resource/Porto>> <<http://dbpedia.org/property/name>> "Porto"@en .
 - <<http://dbpedia.org/resource/Porto>> <<http://xmlns.com/foaf/0.1/nick>> "(\\"The Undefeated City\\"),"@en
 - <<http://dbpedia.org/resource/Porto>> <<http://www.w3.org/2002/07/owl#sameAs>> <<http://www.wikidata.org/entity/Q36433>>
- From the perspective of constructing entity representations, this is conceptually no different from having data stored in relational databases, where the same information is available through fields and foreign-key relationships.

Entity Representations from Triples

- The number of potential fields is huge (in the 1000s), thus we need a different strategy.
- A common solution is predicate folding – grouping predicates together into a small set of predefined categories, so that we have a handful go fields.
- Predicates may be grouped based on their type or importance.
- Common entity fields: name, name variants, attributed, types, outgoing relations, incoming relations, top predicates, catch-all.

RDF Graph

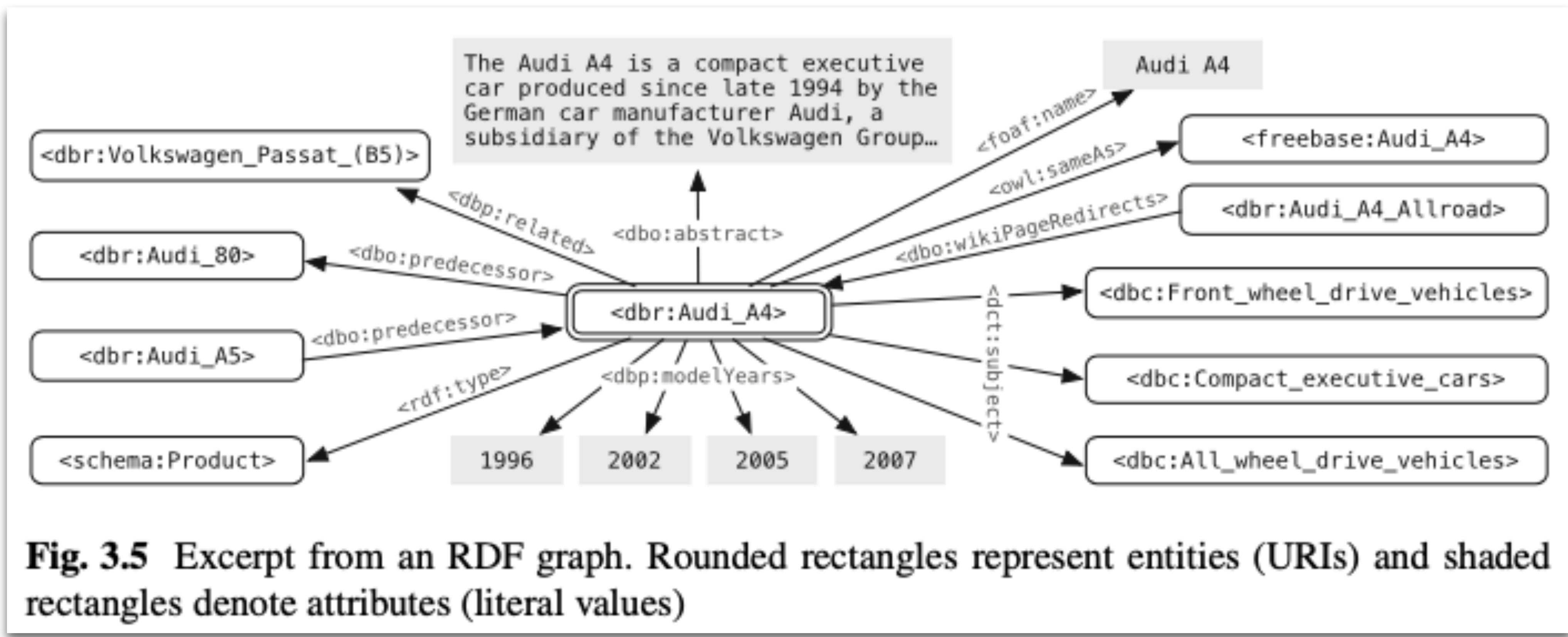


Fig. 3.5 Excerpt from an RDF graph. Rounded rectangles represent entities (URIs) and shaded rectangles denote attributes (literal values)

Fielded Entity Description

Table 3.4 Excerpt from the fielded entity description of AUDI A4, corresponding to Fig. 3.5

Field	Content
Name	Audi A4
Name variants	Audi A4 ... Audi A4 Allroad
Attributes	The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group [...] ... 1996 ... 2002 ... 2005 ... 2007
Types	Product ... Front wheel drive vehicles ... Compact executive cars ... All wheel drive vehicles
Outgoing relations	Volkswagen Passat (B5) ... Audi 80
Incoming relations	Audi A5
<foaf:name>	Audi A4
<dbo:abstract>	The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group [...]
Catch-all	Audi A4 ... Audi A4 ... Audi A4 Allroad ... The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group [...] ... 1996 ... 2002 ... 2005 ... 2007 ... Product ... Front wheel drive vehicles ... Compact executive cars ... All wheel drive vehicles ... Volkswagen Passat (B5) ... Audi 80 ... Audi A5

Commonly Used Fields (1)

- **Name** contains the name(s) of the entity.
 - The two main predicates mapped to this field are <foaf:name> and <rdfs:label>.
 - One might follow a simple heuristic and additionally consider all predicates ending with "name", "label", or "title".
- **Name variants** (aliases) may be aggregated in a separate field.
 - In DBpedia, such variants may be collected via Wikipedia redirects (via <dbo:wikiPageRedirects>) and disambiguations (using <dbo:wikiPageDisambiguates>).
- **Attributes** includes all objects with literal values, except the ones already included in the name field.
 - In some cases, the name of the predicate may also be included along with the value, e.g., "founding date 1964" (vs. just the value part, "1964").

Commonly Used Fields (2)

- **Types** holds all types (categories, classes, etc.) to which the entity is assigned.
 - Commonly, <rdf:type> is used for types;
 - In DBpedia, <dct:subject> is used for assigning Wikipedia categories, which may also be considered as entity types.
- **Outgoing relations** contains all URI objects, i.e., names of entities (or resources in general) that the subject entity links to.
 - If the types or name variants fields are used, then those predicates are excluded;
 - Values might be prefixed with the predicate name, e.g., "spouse Michelle Obama".
- **Incoming relations** is made up of subject URIs from all SPO triples where the entity appears as object.
- **Top predicates** may be considered as individual fields
 - e.g., include the top-1000 most frequent DBpedia predicates as fields.
- **Catch-all** (or content) is a field that amasses all textual content related to the entity.

Triples to Text

- How to represent triples as text?
- Triple objects values are either URIs (links) or literals.
- Literals can be treated as regular text.
- URIs need to be converted to human-readable
 - Some are user-friendly: <https://dbpedia.org/page/Porto>
 - Others are not: <http://www.wikidata.org/entity/Q36433>
- URI resolution is the process of finding the a human-readable name/label for a URI.

URI Resolution

- The specific predicate that holds the name of a resource depends on the RDF vocabulary used.
- Commonly, <foaf:name> or <rdfs:label> are used
- Given a SPO triple, for example
 - <dbr:Audi_A4> <rdf:type> <dbo:MeanOfTransportation>
- The corresponding resources's name is contained in the object element of this triple:
 - <dbo:MeanOfTransportation> <rdfs:label> “mean of transportation”

Ranking Entities

- With the term-based entity representations, we now need to rank entities with respect to their relevance to a search query.
- This can be viewed as a the problem of assigning a score to each entity in the catalog.
- The retrieval models from document ranking — e.g. vector space model, language models, probabilistic models — are used for entity scoring by replacing the *document* with *entity* in the equations.

Entity Ranking Evaluation

- Evaluation of ad hoc entity retrieval is analogous to that of ad hoc document retrieval.
- Given a ranked list of items, the relevance of each item is judged with respect to the underlying information need.
 - Set-based measures, e.g. precision, recall, precision at rank cutoff k (P@10, P@20).
 - Rank-based measures, e.g. average precision (AvP), mean average precision (MAP).
- The use of text collections is the de facto standard standard for evaluation in IR. Some relevant collections that can be used for entity retrieval are: INEX Entity Ranking (XER), TREC Entity, Semantic Search Challenge, INEX Linked Data.

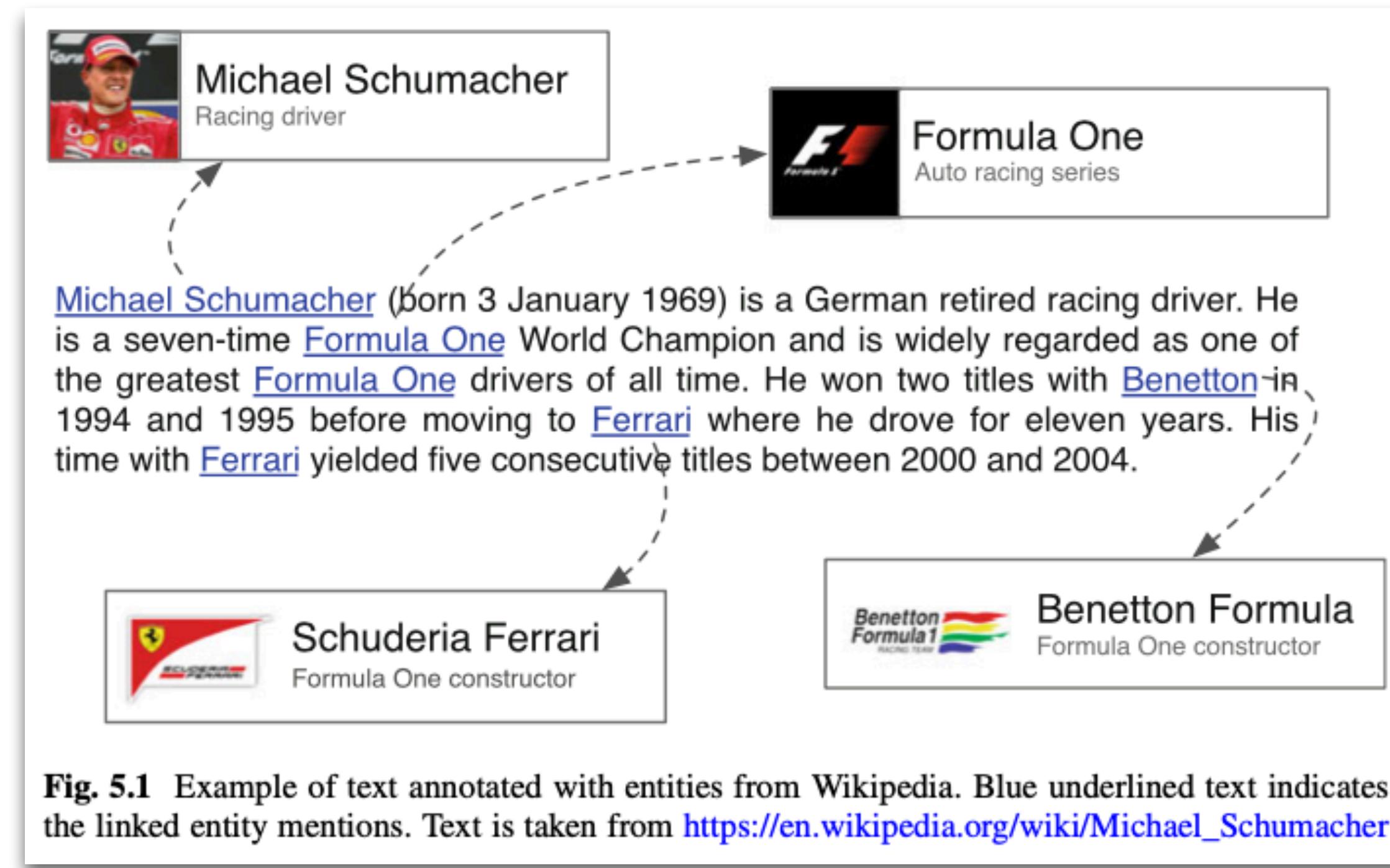
Entity Ranking Summary

- In this section we discussed approaches for ranking entities in various datasets, from unstructured documents, to structured knowledge bases.
- Most of the effort in this area has been in constructing term-based representations of entities, which can be ranked using traditional document retrieval techniques.
- Unstructured entity representations with a bag-of-words retrieval models usually provide solid performance and a good starting point.
- Other approaches exist, most notably the use of specific characteristics of entities, such as relationships.

Entity Linking

Entity Linking

- Entity linking is the task of recognizing entity mentions in text and linking them to the corresponding entries in a knowledge base.



Practical Example (Text)

The screenshot shows the DBpedia Spotlight demo interface. At the top is the logo featuring a stylized yellow flower or tree above the text "DBpedia Spotlight". Below the logo are several input fields and controls:

- Confidence:** A slider set to 0.5.
- Language:** A dropdown menu set to "Portuguese".
- n-best candidates:** A checkbox that is unchecked.
- SELECT TYPES...** and **ANNOTATE** buttons.

In the main text area, there is a block of Portuguese text about Berlin:

Berlim é a capital e um dos dezesseis estados da Alemanha. Com uma população de 3,5 milhões dentro de limites da cidade, é a maior cidade do país, e a sétima área urbana mais povoada da União Europeia. Situada no nordeste da Alemanha, é o centro da área metropolitana de Berlim-Brandemburgo, que inclui 5 milhões de pessoas de mais de 190 nações. Localizada na grande planície europeia, Berlim é influenciada por um clima temperado sazonal. Cerca de um terço da área da cidade é composta por florestas, parques, jardins, rios e lagos. Documentada pela primeira vez no século XIII, Berlim foi sucessivamente a capital do Reino da Prússia (1701-1918), do Império Alemão (1871-1918), da República de Weimar (1919-1933) e do Terceiro Reich (1933-1945). Após a Segunda Guerra Mundial, a cidade foi dividida; Berlim Oriental se tornou a capital da Alemanha Oriental, enquanto Berlim Ocidental se tornou um exclave da Alemanha Ocidental,

At the bottom of the interface, a note states: "This demo uses the statistical [DBpedia-Spotlight](#) web service at <https://api.dbpedia-spotlight.org/pt>".

Practical Example (Annotated)



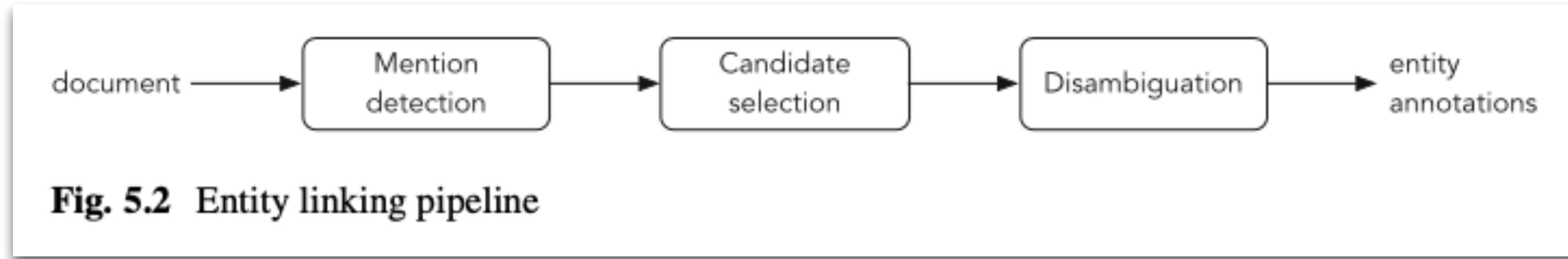
The screenshot shows the DBpedia Spotlight web interface. At the top, there's a logo featuring a stylized blue and yellow flower-like icon above the text "DBpedia Spotlight". Below the logo are several input fields: "Confidence:" with a slider set to 0.5, "Language:" set to "Portuguese", and a checkbox for "n-best candidates" which is unchecked. There are also "SELECT TYPES..." and "ANNOTATE" buttons. The main content area contains a paragraph of text in Portuguese about Berlin, with many words underlined as links to DBpedia entities. At the bottom right of this area is a "BACK TO TEXT" button.

[Berlim](#) é a capital e um dos dezesseis estados da [Alemanha](#). Com uma população de 3,5 milhões dentro de limites da cidade, é a maior cidade do país, e a sétima área urbana mais povoada da [União Europeia](#). Situada no nordeste da [Alemanha](#), é o centro da [área metropolitana](#) de Berlim-Brandemburgo, que inclui 5 milhões de pessoas de mais de 190 nações. Localizada na [grande planície europeia](#), [Berlim](#) é influenciada por um [clima temperado](#) sazonal. Cerca de um terço da área da cidade é composta por florestas, parques, jardins, rios e lagos. Documentada pela primeira vez no [século XIII](#), [Berlim](#) foi sucessivamente a capital do Reino da [Prússia](#) (1701-1918), do [Império Alemão](#) (1871-1918), da [República de Weimar](#) (1919-1933) e do [Terceiro Reich](#) (1933-1945). Após a [Segunda Guerra Mundial](#), a cidade foi dividida; [Berlim Oriental](#) se tornou a capital da [Alemanha Oriental](#), enquanto [Berlim Ocidental](#) se tornou um [exclave](#) da [Alemanha Ocidental](#), cercada pelo [muro de Berlim](#), entre os anos de 1961-1989; a cidade de Bonn tornou-se a capital da [Alemanha Ocidental](#). Após a [reunificação alemã](#) em 1990, a cidade recuperou o seu estatuto como a capital da [República Federal da Alemanha](#), sediando 147 embaixadas estrangeiras

[BACK TO TEXT](#)

This demo uses the statistical [DBpedia-Spotlight](#) web service at <https://api.dbpedia-spotlight.org/pt>.

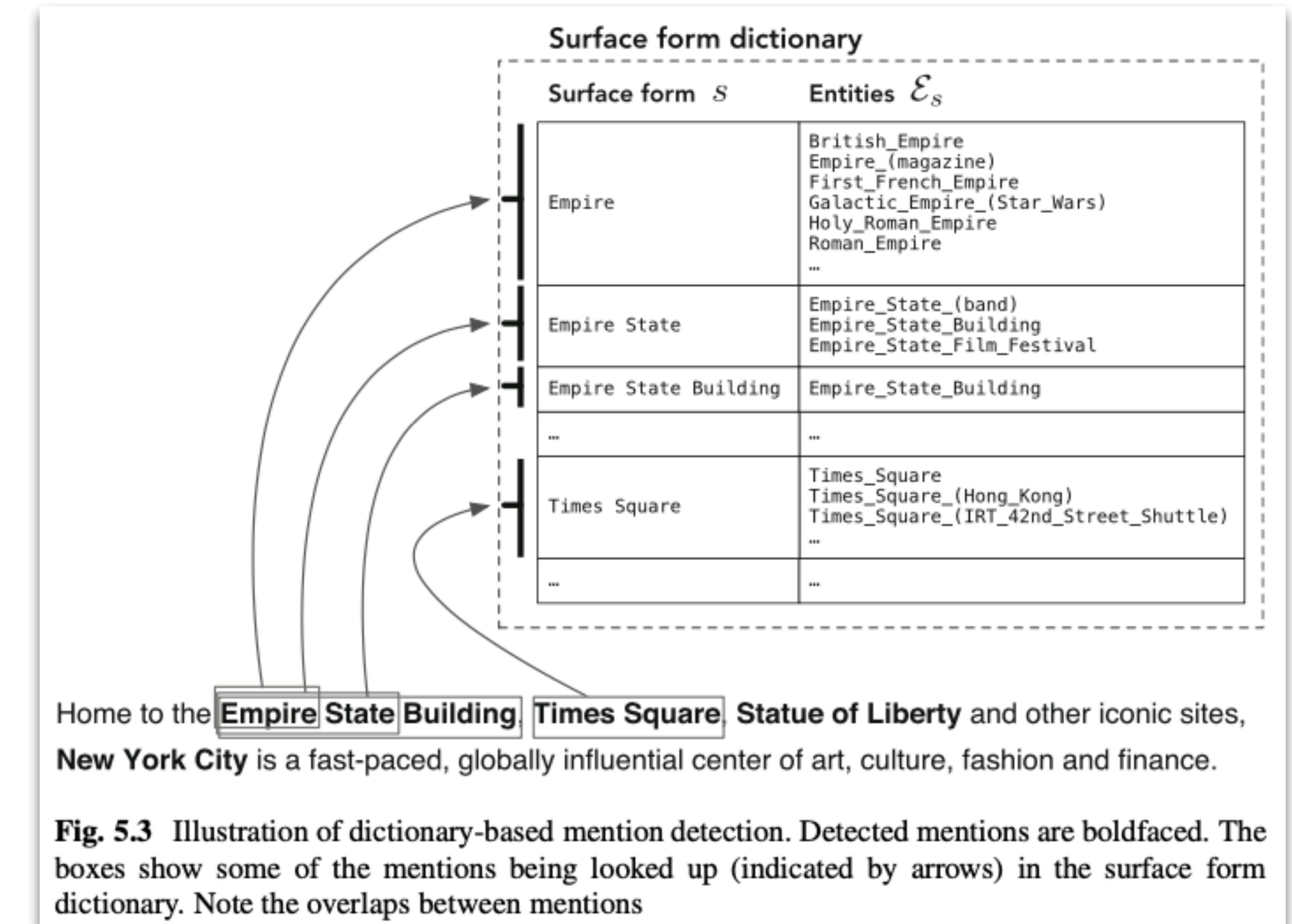
Anatomy of an Entity Linking System



- **Mention detection**, identification of text snippets that can potentially be linked to entities.
- **Candidate selection**, generating a set of candidate entities for each mention.
- **Disambiguation**, selecting a single entity (or none) for each mention, based on the context.

Common Approach for Mention Detection

- Build a dictionary of entity surface forms, i.e. entities with all names variants.
- Check all document n-grams against the dictionary
- Filter out undesired entities



Entity Linking Systems

Table 5.5 Overview of publicly available entity linking systems

System	Reference KR	Online demo	Web API	Source code
AIDA ^a	YAGO2	Yes	Yes	Yes (Java)
DBpedia Spotlight ^b	DBpedia	Yes	Yes	Yes (Java)
Illinois Wikifier ^c	Wikipedia	No	No	Yes (Java)
TAGME ^d	Wikipedia	Yes	Yes	Yes (Java)
Wikipedia Miner ^e	Wikipedia	No	No	Yes (Java)

- <http://www.mpi-inf.mpg.de/yago-naga/aida/>
- <http://spotlight.dbpedia.org/>
- http://cogcomp.cs.illinois.edu/page/download_view/Wikifier
- <https://tagme.d4science.org/tagme/>
- <https://github.com/dnmilne/wikipediaminer>

References and Further Reading

- K. Balog. Entity-Oriented Search. Springer, 2018.
<https://eos-book.org/>
- H. Bast, B. Buchhold, and E. Haussmann. Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval, 10(2–3):119–271, 2016.