Leveraging Knowledge Graphs for Web Search

Part 1 - Introduction to Knowledge Graphs

Gianluca Demartini

University of Sheffield

gianlucademartini.net

Gianluca Demartini

- B.Sc., M.Sc. at U. of Udine, Italy

- Ph.D. at U. of Hannover, Germany

  - Entity Retrieval

g.demartini@sheffield.ac.uk

- Worked at the eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)

- **Lecturer in Data Science** at the iSchool, U. of Sheffield

- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 2013 and ISWC 2013

www.gianlucademartini.net

Research Interests

- **Entity-centric Information Access** (2005-now)
  - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13)
  - NER in Scientific Literature (WWW 14), Prepositions (CIKM 14)
- **Hybrid Human-Machine Systems** (2012-now)
  - ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Memory-based Information Systems (WWW 14, PVLDB)
- **Better Crowdsourcing Platforms** (2013-now)
  - Pick-a-Crowd (WWW 13), Malicious Workers (CHI 15)
  - Scale-up Crowdsourcing (HCOMP 14), Dynamics (WWW 15)

Course Outline

- Part I – **Introduction to Knowledge Graphs**
- Part II – **Named Entity Recognition and Linking to Knowledge Graphs**
- Part III – **Searching for Entities**
- Part IV – **Crowdsourcing for Knowledge Graph Data Quality**

- **Slides here: gianlucademartini.net/kg**

Entity-Centric Information Access

Web Videos Images Shopping News More Search tools

Tom Cruise > Movies
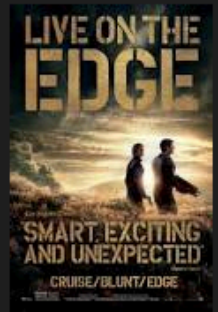
Edge of Tomorrow
2014

Mission: Impossible
1996

Mission: Impossible – Ro...
2015

Top Gun
1986

Mission: Impossible – Gh...
2011

Oblivion
2013

Jack Reacher
2012

Mission: Impossible III
2006

Jerry Maguire
1996

Tom Cruise filmography - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Tom_Cruise**_filmography ▾
Tom Cruise is an American actor and producer who made his film debut with a minor role ... As of 2015, Cruise has reprised his role as Hunt in four more **films** in the .....
"**Movie** Review : Utility Vehicle : '**Days of Thunder**': The NASCAR racing footage and **Tom Cruise's** grin are fine. ... "**The Last Samurai Movie** Review (2003)".
Rock of Ages - Ask the Dust (film) - Losin' It

Tom Cruise - IMDb
www.imdb.com/name/nm0000129/ ▾
Tom Cruise and Arnold Schwarzenegger in Terminator Genisys (2015) **Tom Cruise** at ... Visit our IMDb Picks section to see our recommendations of **movies** and TV ......
movie preview - · **Mission: Impossible - Ghost Protocol** -- An exclusive clip ...

Tom Cruise

Actor

Tom Cruise is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Endless Love. Wikipedia

Born: July 3, 1962 (age 53), Syracuse, New York, United States

Height: 1.70 m

Full name: Thomas Cruise Mapother IV

6

- Entity-seeking queries make up 40-50% of the query volume
  - Jeffrey Pound, Peter Mika, Hugo Zaragoza: Ad-hoc object retrieval in the web of data. WWW 2010: 771-780
  - Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, Ariel Fuxman: Active objects: actions for entity-centric search. WWW 2012: 589-598

- Show a summary of the most likely information-needs
  - Including related entities for navigation
  - *Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, Nicolas Torzec: Entity Recommendations in Web Search. ISWC 2013*

Matt Damon
Actor

© Getty Images

Matthew Paige "Matt" Damon is an American actor, voice actor, screenwriter, producer, and philanthropist whose career was launched following the success of the drama film Good Will Hunting (1997) from a screenplay... wikipedia.org

Born: October 8, 1970 (age 43), Cambridge, Massachusetts, USA

Height: 5' 10" (1.78m)

Spouse: Luciana Barroso (m. 2005-present)

Partner: Winona Ryder (1998-2000)

Parents: Kent Damon, Nancy Carlsson-Paige

Children: Isabella Damon, Alexia Barroso, Gia Zavala Damon, Stella Damon

Movies & TV Shows

The Zero Theorem   Elysium   The Departed   We Bought a Zoo   Good Will Hunting
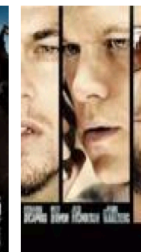
Feedback

Personalization in on-line news

- Entity linking
- Entity ranking based on the relevance to the document
- Entity type ranking
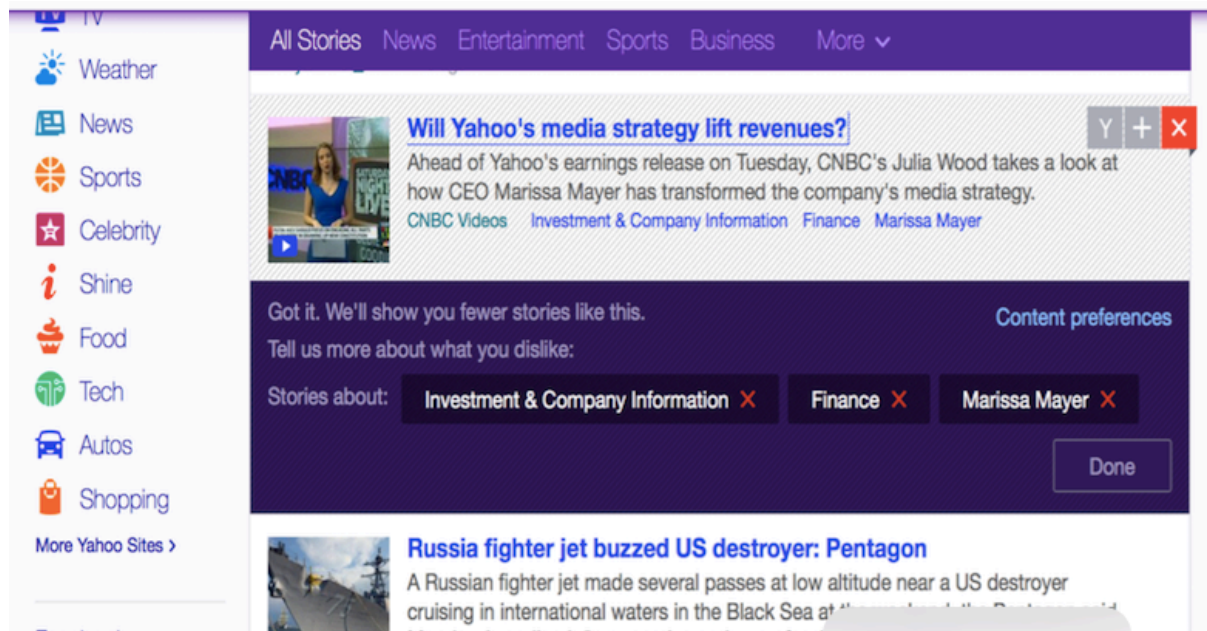
Natural language interfaces
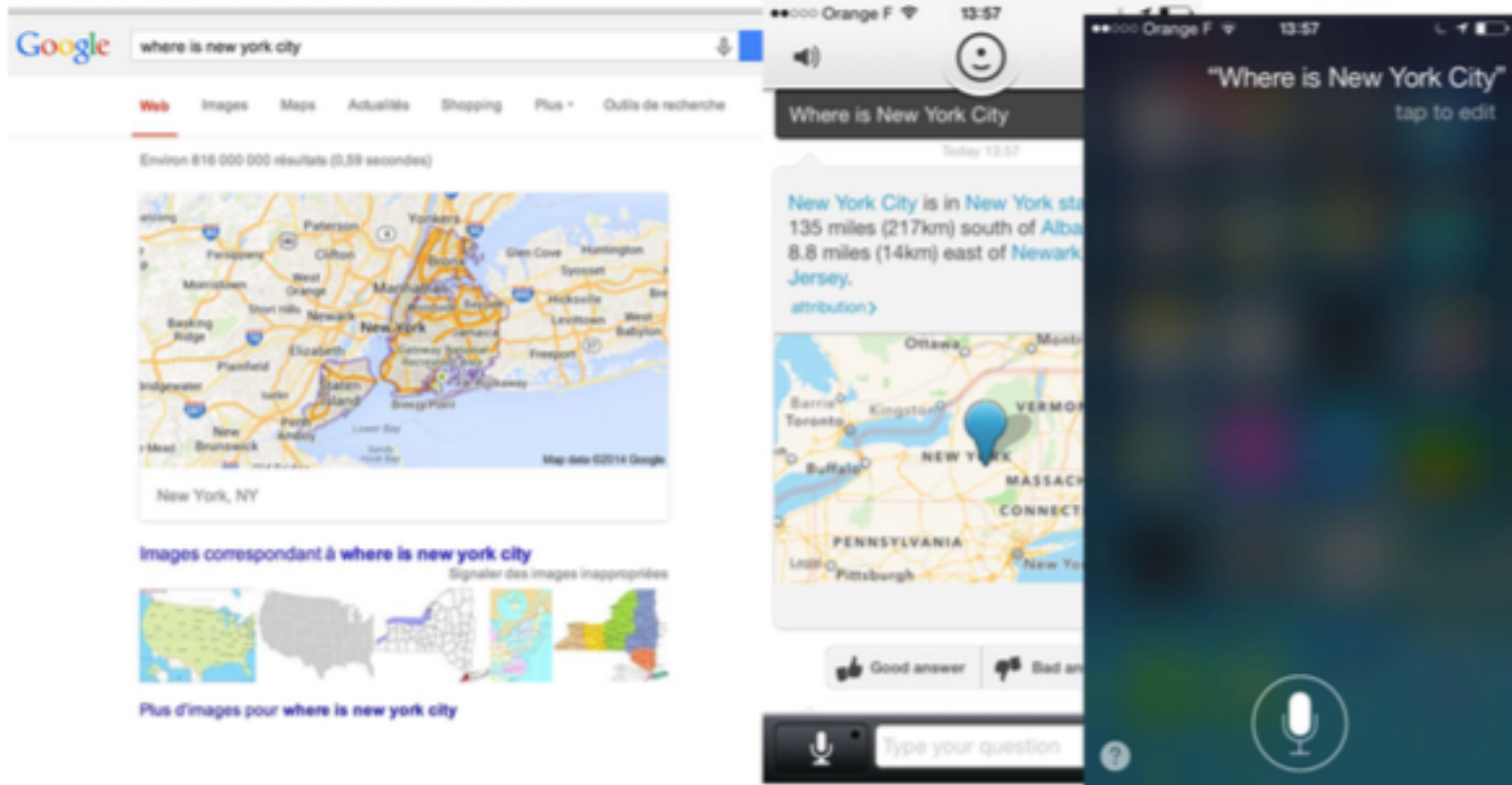
Mobile search

- Information access on-the-go requires hands-free operation
  - Driving, walking, gym, etc.
    - Americans spend 540 hours a year in their cars [1] vs. 348 hours browsing the Web [2]
- ~50% of queries are coming from mobile devices (and growing)
  - Changing habits, e.g. iPad usage peaks before bedtime
  - Limitations in input/output

[1] http://answers.google.com/answers/threadview?id=392456
[2] http://articles.latimes.com/2012/jun/22/business/la-fi-tn-top-us-brands-news-web-sites-20120622

Mobile search challenges and opportunities

- Interaction
  - Question-answering
  - Support for interactive retrieval
  - Spoken-language access
  - Task completion
- Contextualization
  - Personalization
  - Geo
  - Context (work/home/travel)
    - Try getaviate.com

Task completion

- Help users in task completion
  - But users have been trained to talk in nouns
    - Retrieval performance **decreases** by adding verbs to queries
  - We need to understand what the available actions are

- Modeling actions
  - Understand what **actions** can be taken on a page
  - Help users in mapping their query to potential actions
  - Applications in web search, email etc.

Applications

Email (Gmail)

SERP (Yandex)

Latest academic event

- The SIGIR2014 Entity Recognition and Disambiguation Challenge on linking web search queries to a KG

- http://web-ngram.research.microsoft.com/ERD2014/

Web of Data

- Freebase
  - Acquired by Google in July 2010.
  - Knowledge Graph launched in May 2012.
  - Read-only in December 2014 -> WikiData

- Schema.org
  - Driven by major search engine companies
  - Machine-readable annotations of Web pages

- Linked Open Data
  - 31 billion triples, Sept 2011
  - 90 billion triples, Aug 2015 (stats.lod2.eu)
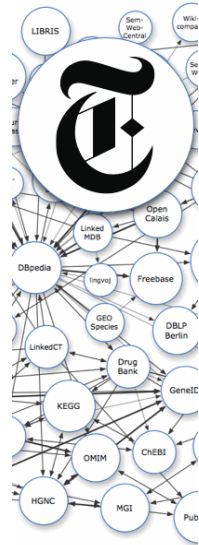
Schema.org

- Agreement between Bing, Google, Yahoo, and Yandex on what markup webmasters should use
  - Help adoption by reducing fragmentation
  - Pre-competitive: each party will continue to build competing products independently
- Schema.org covers areas of interest to all three parties
  - Business listings (local), creative works (video), recipes, reviews

Example: schema.org

Thing > Organization > LocalBusiness > FoodEstablishment > Restaurant
A restaurant.

Property	Expected Type	Description
Properties from Thing		
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from Place		
address	PostalAddress	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
containedIn	Place	The basic containment relation between places.
events	Event	Upcoming or past events associated with this place or organization.
faxNumber	Text	The fax number.
geo	GeoCoordinates or GeoShape	The geo coordinates of the place.
interactionCount	Text	A count of a specific user interactions with this item—for example, 20 UserLikes, 5 UserComments, or 300 UserDownloads. The user interaction type should be one of the sub types of UserInteraction.
maps	URL	A URL to a map of the place.
photos	Photograph or	Photographs of this place.

Linked Open Data

2007

Z. Kaoudi and I. Manolescu, ICDE seminar 2013
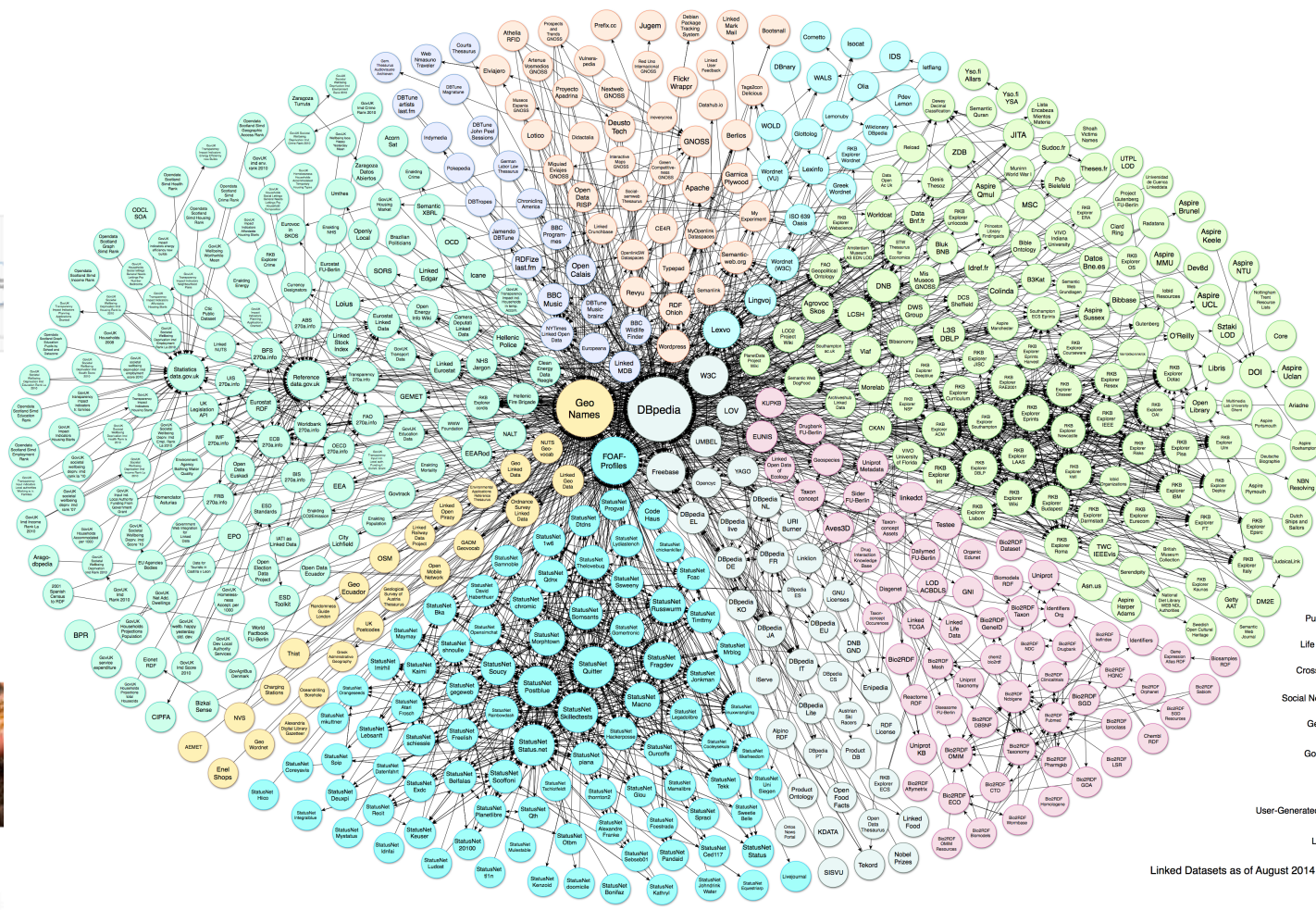
Popular Knowledge Graphs

- Public KGs
  - Freebase: 50M entities
  - Wikidata: 14M entities
  - Extracted from Wikipedia
    - DBPedia: 4.5M entities
    - YAGO: 10M entities (multilingual)
- Corporate KGs
  - Google KG -> Knowledge Vault
  - Facebook Entity Graph
  - Microsoft Satori (mid-2013)

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
Knowledge Vault (KV)	1100	45M	4469	271M
DeepDive [32]	4	2.7M	34	7M^a
NELL [8]	271	5.19M	306	0.435M^b
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M^c
Freebase [4]	1,500	40M	35,000	637M^d
Knowledge Graph (KG)	1,500	570M	35,000	18,000M^e

Wikipedia

Paul Allen

From Wikipedia, the free encyclopedia
(Redirected from Paul Allen's house)

For other people named Paul Allen, see Paul Allen (disambiguation).

Paul Gardner Allen (born January 21, 1953) is an American business magnate and investor. Allen co-founded Microsoft with Bill Gates. He is also the 57th richest along with Viktor Vekselberg and Gerald Cavendish Grosvenor (and family) who all rank the same with an estimated wealth of $13 billion as of 2011. He is the founder and chairman of Vulcan Inc., which manages his business and philanthropic efforts. Allen also has a multi-billion dollar investment portfolio which includes technology companies, real estate holdings, and stakes in other technology, media, and content companies. Allen also owns two professional sports teams, the Seattle Seahawks of the National Football League (NFL),[2] and the Portland Trail Blazers of the National Basketball Association (NBA).[3] He is also part-owner of the Seattle Sounders FC, which joined Major League Soccer (MLS) in 2009.[4] Allen's memoir *Idea Man: A Memoir by the Cofounder of Microsoft* was released on April 19, 2011.

Paul Allen

Contents [hide]

1 Life and career
    1.1 Microsoft
    1.2 Recognition
2 Philanthropy

Categories: 1953 births | Living people | American billionaires | American computer businesspeople | American computer programmers | American philanthropists | American real estate businesspeople | American soccer chairmen and investors | American technology company founders | Businesspeople in software | History of Microsoft | Major League Soccer executives | Microsoft employees | National Basketball Association owners | National Football League owners | People from Seattle, Washington | People with cancer | Portland Trail Blazers owners | Seattle Seahawks owners | Women's National Basketball Association executives

|	
orn	January 21, 1953 (age 59) Seattle, Washington, U.S.
esidence	Mercer Island, Washington, U.S.
ccupation	Chairman of Vulcan Inc.
et worth	▲ US$14.4 billion (2012)[1]

20

Dbpedia

dbpedia-owl:birthDate	▪ 1953-01-21 (xsd:date)
dbpedia-owl:birthYear	▪ 1953-01-01 00:00:00 (xsd:date)
dbpedia-owl:networth	▪ 1.3E10
dbpedia-owl:occupation	▪ dbpedia:Vulcan_Inc. ▪ dbpedia:Paul_Allen__1
dbpedia-owl:residence	▪ dbpedia:Mercer_Island,_Washington
dbpedia-owl:thumbnail	▪ http://upload.wikimedia.org/wikipedia/commons/thumb/5/51/Paull_Allen_fix_1.JPG/200px-Paull_Allen_fix_
dbpedia-owl:wikiPageExternalLink	▪ http://books.google.com/books?id=iBhNuAj_Tx4C&printsec=frontcover ▪ http://money.cnn.com/magazines/fortune/fortune_archive/1995/10/02/206528/index.htm ▪ http://www.thocp.net/biographies/allen_paul.htm ▪ http://www.paulallen.com/ ▪ http://www.oregonencyclopedia.org/entry/view/allen_paul_1953_/ ▪ http://www.forbes.com/profile/paul-allen ▪ http://books.google.com/books?id=3lFczEsFLRsC&printsec=frontcover ▪ http://seahawknationblog.com/2011/01/seattle-seahawks-owner-paul-allen-is-sports-citizen-of-the-year-2/%
dbpprop:alternativeNames	▪ Allen, Paul Gardner
dbpprop:before	▪ dbpedia:Ken_Behring ▪ dbpedia:Larry_Weinberg
dbpprop:birthDate	▪ 1953-01-21 (xsd:date)
dbpprop:birthPlace	▪ Seattle
dbpprop:dateOfBirth	▪ 1953-01-21 (xsd:date)
dbpprop:name	▪ Allen, Paul ▪ Paul Gardner Allen
dbpprop:networth	▪ US$13 billion
dbpprop:occupation	▪ Chairman, Vulcan Inc.
dbpprop:placeOfBirth	▪ Seattle, U.S.
dbpprop:residence	▪ dbpedia:Mercer_Island,_Washington
dbpprop:shortDescription	▪ Co-Founder of Microsoft
dbpprop:title	▪ Seattle Sounders FC owner ▪ Portland Trail Blazers owner ▪ Seattle Seahawks owner
dbpprop:wikiPageUsesTemplate	▪ dbpedia:Template:Infobox_person ▪ dbpedia:Template:Persondata ▪ dbpedia:Template:Succession_box

Using the Dbpedia ontology

Raw data

21

Problem definition

- Information extraction + knowledge acquisition

knowledge graph

extracted infomation

(textual) data

Information extraction

Knowledge acquisition
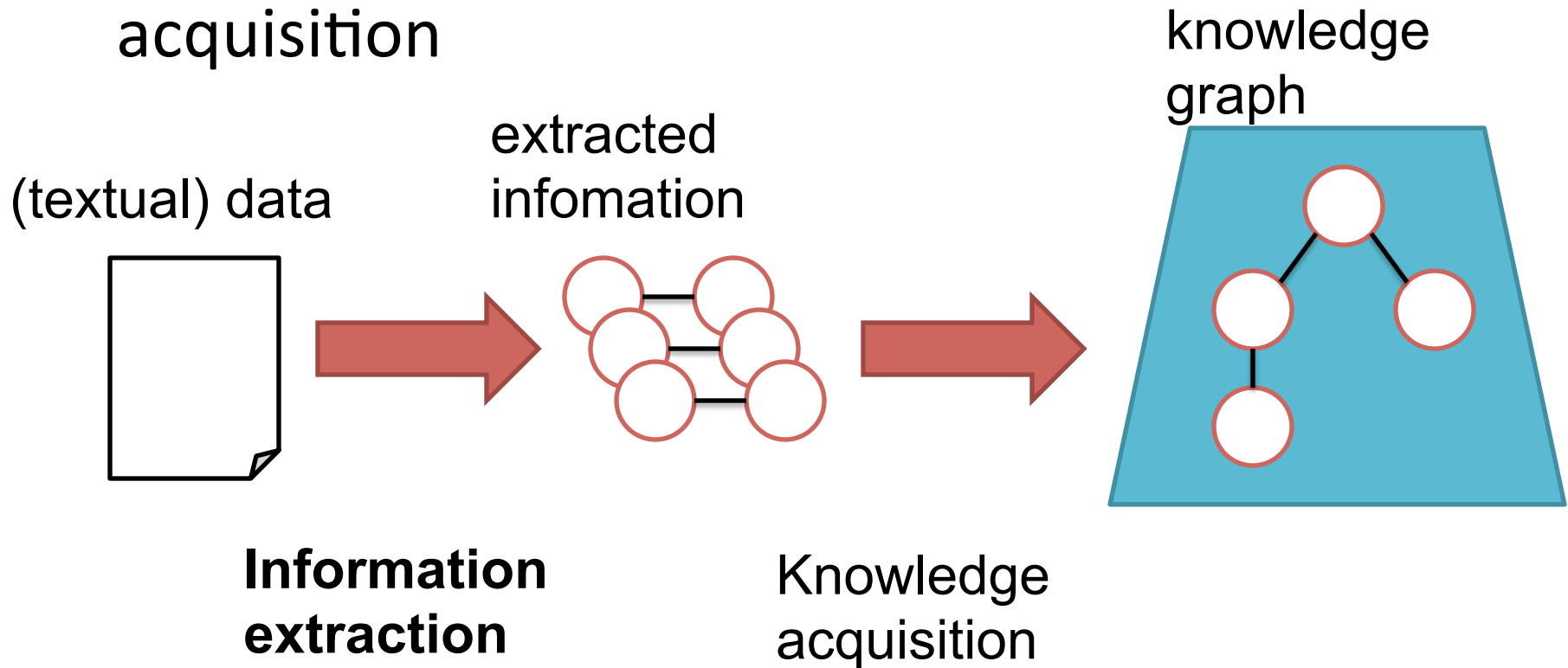
Information extraction

- From the advent of the WWW, there are **huge quantities of unstructured textual data**, where manual information extraction would be infeasible
- How to extract information from text automatically with human-comparable quality

Information extraction: early solutions

- Match manually defined **patterns** against text
- Example:
  - Patterns like *"Pay ? from ? in favor of ?"*
  - ATRANS (1986) inter-banking message exchange

Knowledge acquisition

- Constructing a knowledge base is expensive
  - The Cyc KB was mostly manually constructed over the last 20 years (250K entities)

- Coupling information extraction and knowledge acquistion lets us construct a knowledge base with no or **little human effort**

Human effort

- Challenges:
  - Defining (domain-specific and domain-independent) extraction patterns
  - Especially, in case of bootstrapping approaches:
    - Specifying relations (x „is the capital city of" y)
    - Construction of training examples (Berlin, Germany)
  - Maintaining knowledge base consistency
    - "born in 1946", "born in 1950"
- Crowdsourcing?

Constructing Knowledge Graphs

- Manually-supported construction
  - Freebase
  - WikiData
- Automatic methods for constructing knowledge bases
  - Extracted from semi-structured data: DBPedia, YAGO
  - Text-based extraction
  - Text+DB: Dong et al., "Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion", KDD 2014

Some challenges when creating KGs

- Which entities to include?
  - Notable entities vs *tail* entities
- Statement validity
  - Update vs timestamp
  - For example, married_with annotated with "from" "to" year

General knowledge extraction tools

- WebKB

- TextRunner http://openie.cs.washington.edu/

- Cyc

- SOFIE with the corresponding YAGO knowledge base

- Read The Web

- EntityCube
  http://entitycube.research.microsoft.com/

Data Formats

Resource Description Framework

- RDF, building block of the Semantic Web

- Used to encode data as *triples,* forming *distributed graphs*

- Standardized by the W3C
  - http://www.w3.org/RDF/
  - About a dozen recommendations (i.e., standards) http://www.w3.org/standards/techs/rdf#w3c_all

RDF Triples

- Used to encode data in a semi-structured way

- Like small sentences: subject + attribute-value

1:subject, 2:predicate, 3:object

ex.: gianluca made zencrowd_paper

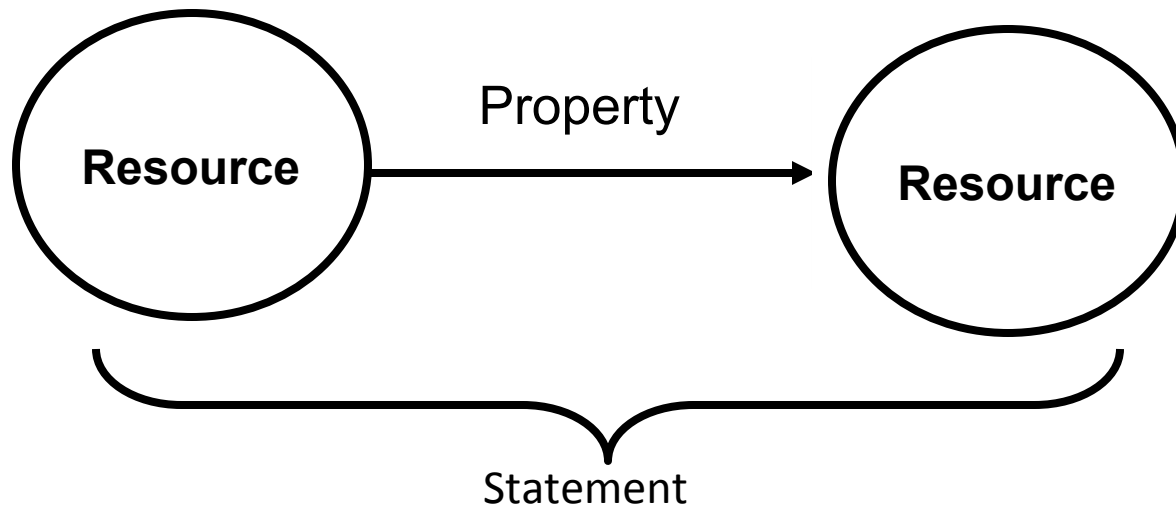1: http://data.semanticweb.org/person/gianluca-demartini

2: http://xmlns.com/foaf/0.1/made

3: http://data.semanticweb.org/conference/www/2012/paper/982

Subject: URI

Predicate: URI

Object: URI / value ("literal")

RDF Model

Resource —Property→ Resource

Statement

Naturally Forms Distributed Graphs

db:Massachusetts

dbo:capital

db:Boston

dbo:inState   dbo:nickname

dbo:governor

db:Deval_Patrick

dbo:nickname

dbo:population

The Bay State

dbo:population

dbo:friend

Beantown

2,320,400
xsd:integer

db:Barack_Obama

dbo:livesIn

642,109
xsd:integer

FollowsOnTwitter

FollowsOnTwitter

<http://www.w3.org/2010/Talkes/1107-rdf-sandro/slides>

dc:reator

<http://www.w3.org/
People/Sandro/data
#Sandro_Hawke>

```
db:Boston dbo:nickname "Beantown".
db:Boston dbo:population "642109"^^xsd:integer.
db:Boston dbo:inState db:Massachusetts.
db:Massachusetts dbo:capital db:Boston.
db:Massachusetts dbo:nickname "The Bay State".
```

Graphs © Sandro Hawke, W3C

34

13   UNIVERSITAS
      FRIBURGENSIS

XML Serialization Example


```
<RDF xmlns = "http://www.w3.org/TR/WD-rdf-syntax#"
        xmlns:dc = "http://purl.org/dc/elements/1.0/">
  <Description about = "URI:R">
    <dc:Title> RDF Presentation </dc:Title>
    <dc:Creator> Paul Miller </dc:Creator>
  </Description>
</RDF>
```

See also JSON-LD

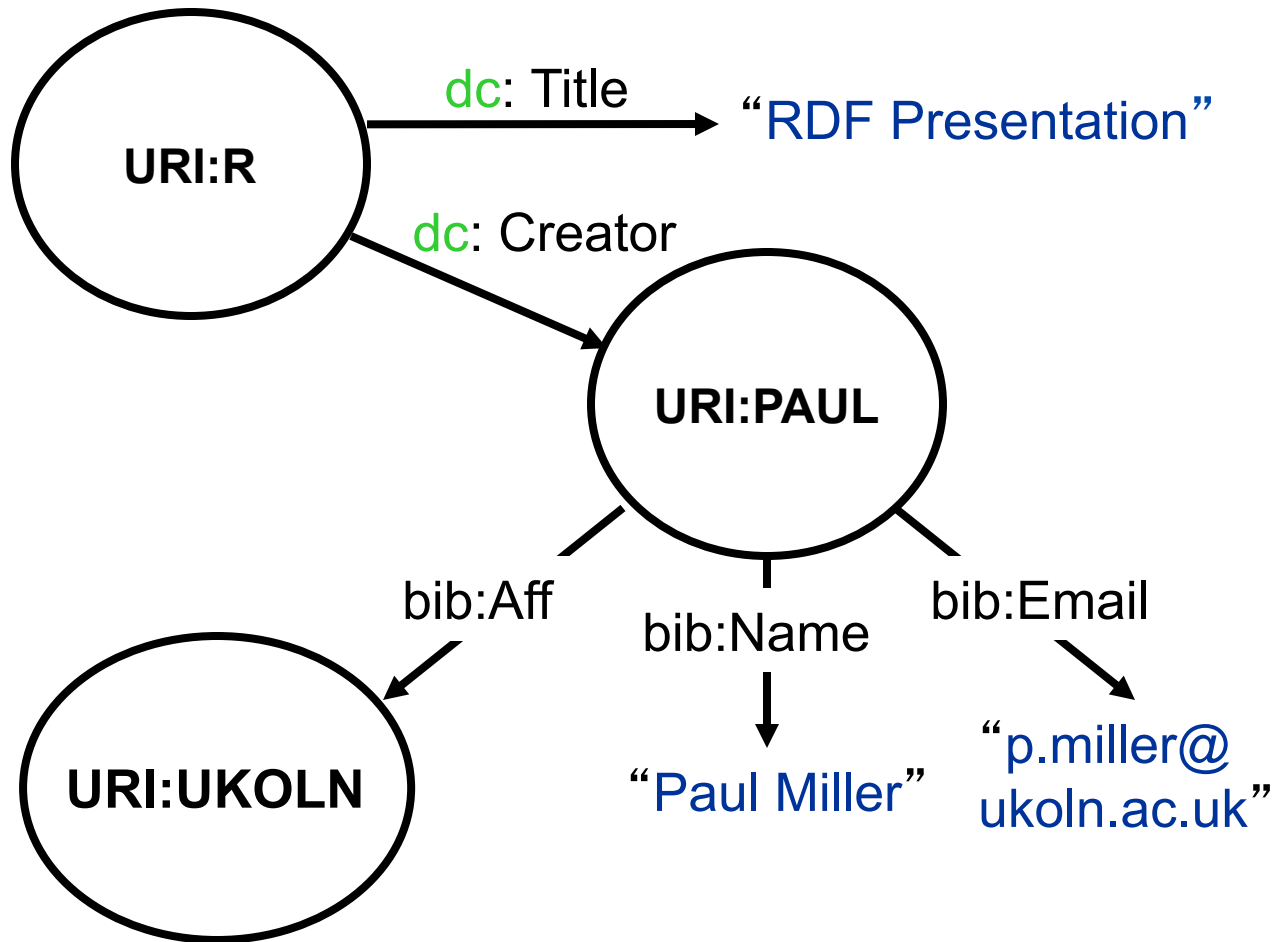# XML Serialization Example #2

XML Serialization Example #2

```
<RDF xmlns = "http://www.w3.org/TR/WD-rdf-syntax#"
        xmlns:dc = "http://purl.org/dc/elements/1.0/"
        xmlns:bib = "http://www.bib.org/persons#">
    <Description about = "URI:R">
      <dc:Title> RDF Presentation </dc:Title>
      <dc:Creator>
        <Description>
          <bib:Name> Paul Miller </bib:Name>
          <bib:Email> p.miller@ukoln.ac.uk </bib:Email>
          <bib:Aff resource = "http://www.ukoln.ac.uk" />
        </Description>
      </dc:Creator>
    </Description>
</RDF>
```

RDF Schemas

- Declaration of vocabularies
  - Properties/Classes defined by a particular community
  - characteristics of properties and/or constraints on corresponding values

- Expressible in the RDF model and syntax
- Provides structure!
 - easier to store, check, and process data

Examples of Vocabularies

- Friend of a Friend (Social Networks)
  - foaf:name
- Dublin Core (Publications)
  - dc:creator, dcterms:temporal
- Good Relations (Products)
  - gr:ProduceOrServiceModel, …

RDF Schema Constructs

- Create classes
  - Class (as in OO, typed instances appearing as subjects/objects)
  - Property (typed predicates)
- Create hierarchies
  - SubClassOf, SubPropertyOf
  - Inheritance mechanisms
- Create constraints on the triples
  - Domain (restricts the subject of a property)
  - Range (restricts the object of a property)

RDF Schema Example
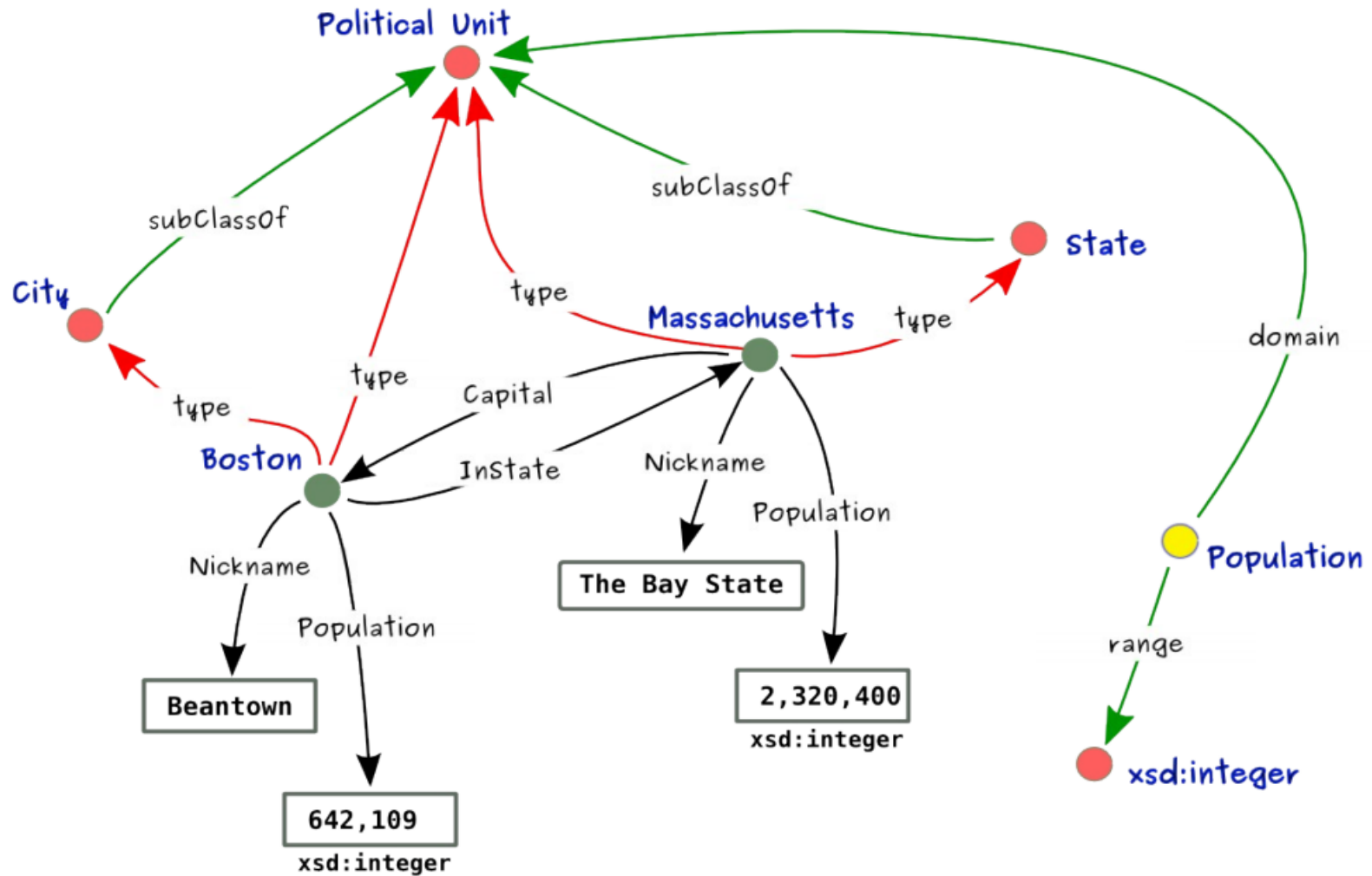
RDFa

- Embedding RDF information in HTML pages
Supported by Google, Yahoo, etc.
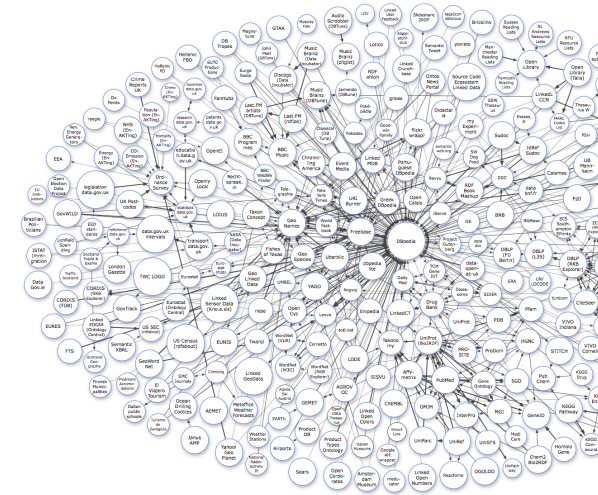
```
<body>
  <div about="http://dbpedia.org/resource/Massachusetts">The
  Massachusetts governor is
    <span rel="db:Governor">
   <span about="http://dbpedia.org/resource/Deval_Patrick">Deval
   Patrick
   </span>,
    </span>
    the nickname is "<span property="db:Nickname">Bay State</span>",
    and the capital
    <span rel="db:Capital">
   <span about="http://dbpedia.org/resource/Boston">
    has the nickname "<span property="db:Nickname">Beantown</span>".
   </span>
    </span>
  </div>
</body>
```

Structured Data about Entities

- Linked Open Data
  - Standard machine-readable formats
  - Available
  - Linked each other

- Embedded RDFa in HTML pages
  - Structured / machine-readable
  - Assigned to a specific webpage
  - More and more popular

OWL

- The Web Ontology Language
  - Very expressive schemas! (ontologies)
  - Description Logics
    - ... and several flavours
  - Example: OWL 2 EL axioms:

    class inclusion (SubClassOf)

    class equivalence (EquivalentClasses)

    class disjointness (DisjointClasses)

    object property inclusion (SubObjectPropertyOf) with or without property chains, and data property inclusion (SubDataPropertyOf)

    property equivalence (EquivalentObjectProperties and EquivalentDataProperties),

    transitive object properties (TransitiveObjectProperty)

    reflexive object properties (ReflexiveObjectProperty)

    domain restrictions (ObjectPropertyDomain and DataPropertyDomain)

    range restrictions (ObjectPropertyRange and DataPropertyRange)

    assertions (SameIndividual, DifferentIndividuals, ClassAssertion, ObjectPropertyAssertion, DataPropertyAssertion, NegativeObjectPropertyAssertion, andNegativeDataPropertyAssertion)

    functional data properties (FunctionalDataProperty)

    keys (HasKey)

  - Inference! ex.: TransitiveObjectProperty(hasAncestor)
    hasAncestor(x, y) $\land$ hasAncestor(y, z) $\rightarrow$ hasAncestor(x, z)

SPARQL

- Declarative query language for RDF/S
  - SQL for the Semantic Web!

```
prefix db: <http://dbpedia.org/resource/>
prefix dbo: <http://dbpedia.org/ontology/>

SELECT ?cap
WHERE { db:Massachusetts dbo:capital ?cap }
```

- Uses *triple patterns*
  - *?subject   ?predicate ?object*

More Complex SPARQL Query

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?title2
WHERE
{
    ?doc       dc:title    "SPARQL at speed" .
    ?doc       dc:creator  ?c .
    ?docOther  dc:creator  ?c .
    ?docOther  dc:title    ?title2

}
```

- On an abstracts/papers database:
  "Find other papers by the authors of a given paper."

Optional graph patterns

Data

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix : <http://example.org/book/> .
@prefix ns: <http://example.org/ns#> .
:book1 dc:title "SPARQL Tutorial" .
:book1 ns:price 42 .
:book2 dc:title "The Semantic Web" .
:book2 ns:price 23 .
```

Query

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ns: <http://example.org/ns#>
SELECT ?title ?price
WHERE { ?x dc:title ?title .
        OPTIONAL { ?x ns:price ?price .
                   FILTER ?price < 30 }}
```

Query Result

title	price
"SPARQL Tutorial"	
"The Semantic Web"	23

SPARQL Constructs

- Many other constructs
  - Order By
  - Distinct
  - Limit
  - Construct
  - Ask
  - Value tests
  - Transitive closures!

Summary

- Semantic Web built on a series of W3C standards
  - URIs (building block of the WWW!)
  - RDF (encode data)
  - RDFS (schemas to define vocabularies and constraint data)
  - SPARQL (query language)

Open Research Problems

- **Growth:** knowledge graphs are incomplete!
  - *Link prediction*: add relations
  - *Ontology matching*: connect graphs
  - *Knowledge extraction*: extract new entities and relations from web/text

- **Validation:** knowledge graphs are not always correct!
  - *Entity resolution*: merge duplicate entities, split wrongly merged ones
  - *Error detection*: remove false assertions

- **Interface:** how to make it easier to access knowledge?
  - *Semantic parsing*: interpret the meaning of queries
  - *Question answering*: compute answers using the knowledge graph

- **Intelligence:** can AI emerge from knowledge graphs?
  - *Automatic reasoning* and planning
  - Generalization and abstraction

by E. Gabrilovich

What's next

- Named Entity Extraction from text
- Entity Linking from text to a KGs
- Selecting relevant facts from the KGs