Technologies du Web
Master COMASIC
Information Extraction and the Semantic Web

Antoine Amarilli[1]

December 2, 2014

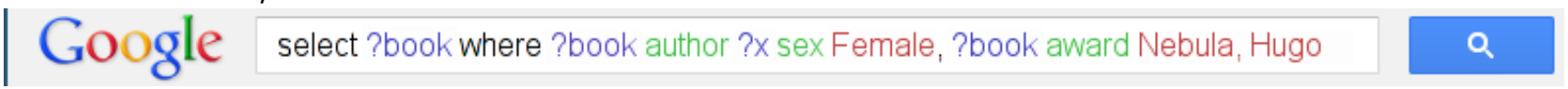
[1] Course material adapted from Fabian Suchanek's slides: `http://suchanek.name/work/teaching/IESW2010.pdf`.
SPARQL example from: `en.wikipedia.org/w/index.php?title=SPARQL&oldid=575552762`.
Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. `http://lod-cloud.net/`

Motivation

- We have seen how search engines work at the level of words.
- Sometimes, this works...

  List of joint winners of the **Hugo and Nebula** awards - Wikipedia, the **...**
  en.wikipedia.org/.../List_of_joint_winners_of_the_**Hugo_and_Nebul**...
  This is a list of the works that have won both the **Hugo Award** and the **Nebula Award**, **awarded** annually to works of science fiction literature. The **Hugo** Awards **...**

- Sometimes, it doesn't:

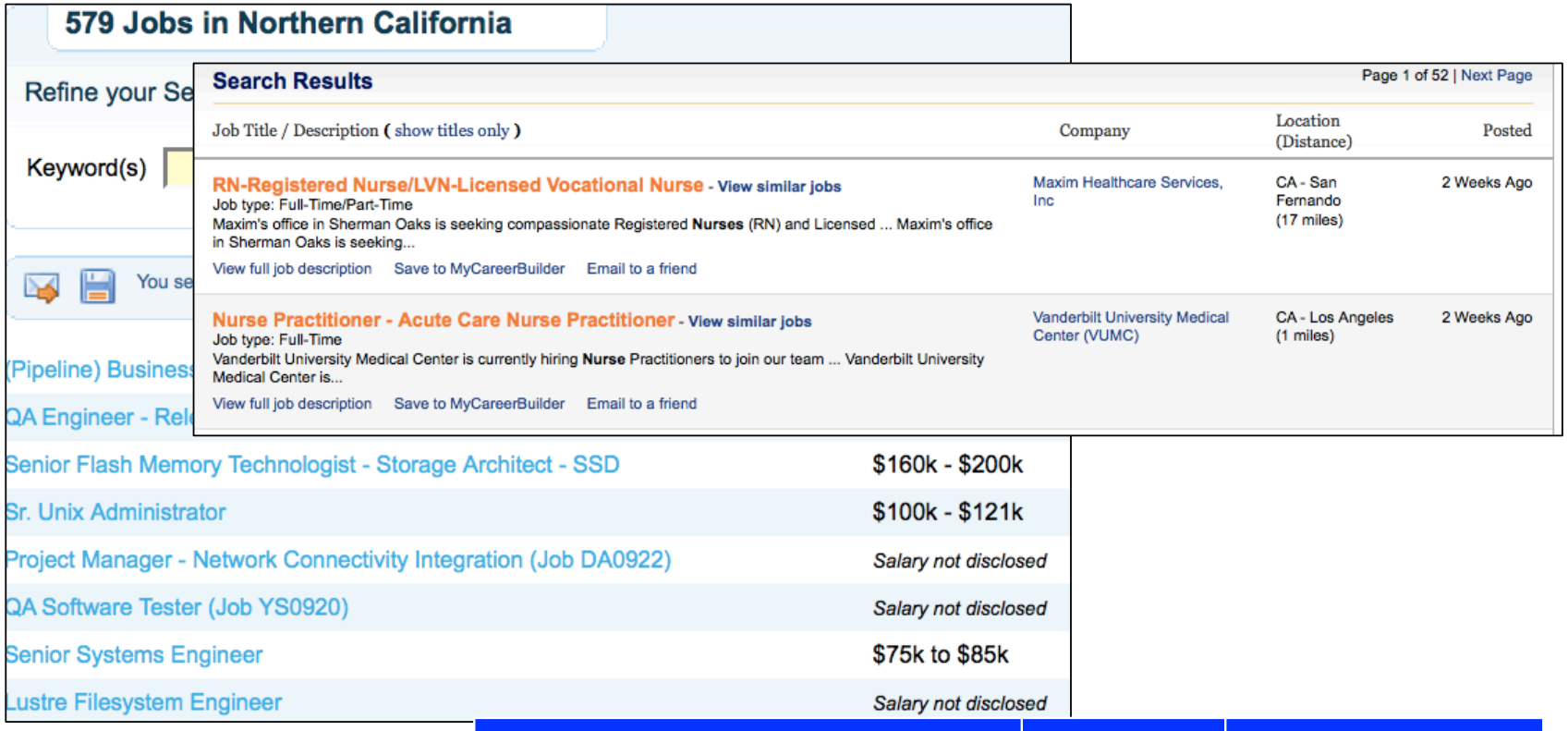
  

- Those hard queries would be easy on RDBMSes!

→ We need to extract structured information.

→ We would like to understand its semantics.

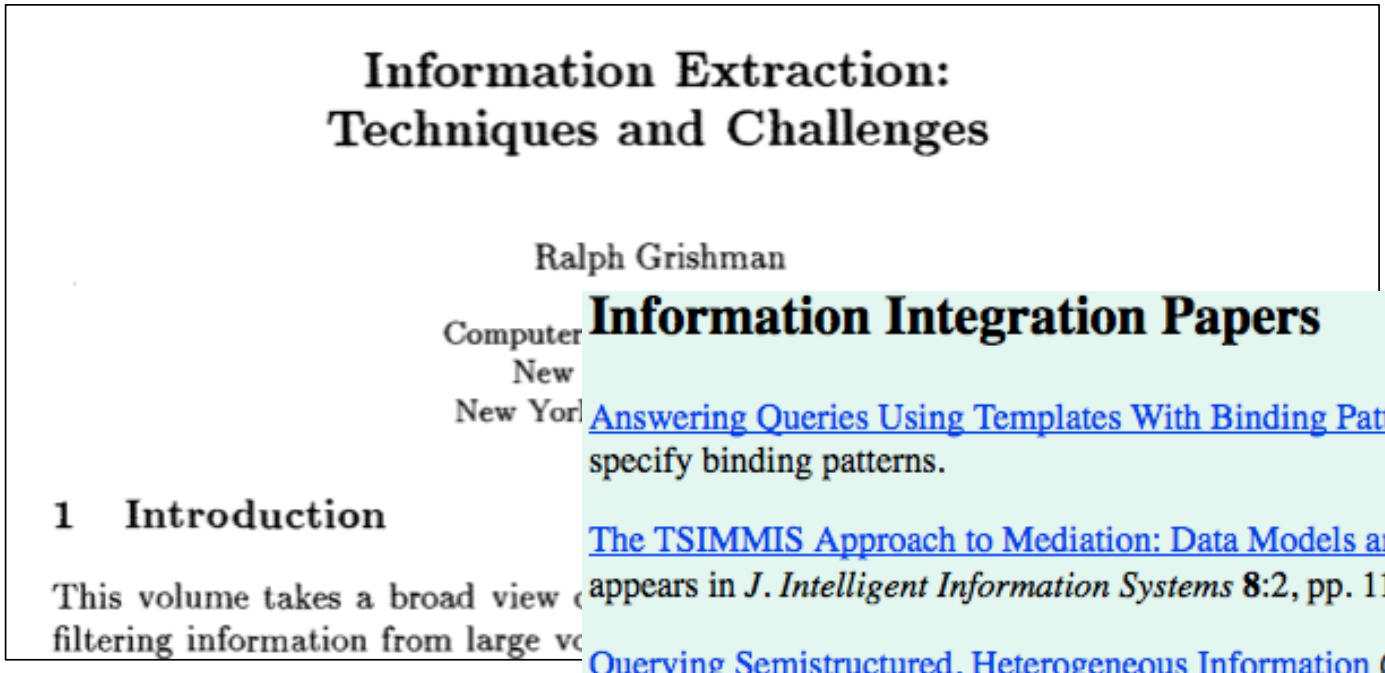
Other motivations: job offerings



Title	Type	Location
Business strategy Associate	Part time	Palo Alto, CA
Registered Nurse	Full time	Los Angeles
...	...	...

8

Other motivations: scientific papers



Information Extraction:
Techniques and Challenges

Ralph Grishman

Computer
New
New Yor

1 Introduction

This volume takes a broad view
filtering information from large v

Information Integration Papers

Answering Queries Using Templates With Binding Patterns. In PODS 1995,
specify binding patterns.

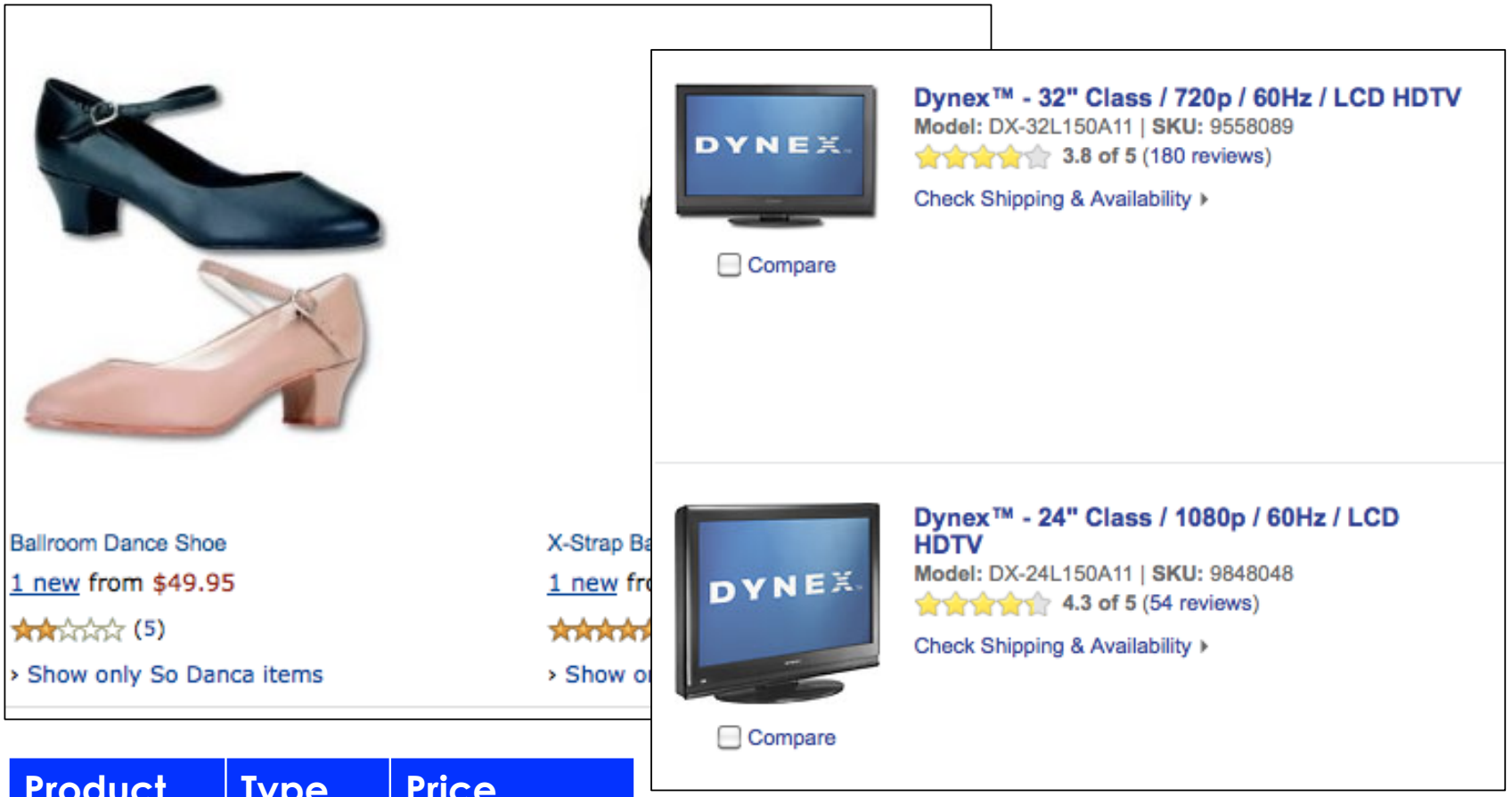
The TSIMMIS Approach to Mediation: Data Models and Languages. A surv
appears in *J. Intelligent Information Systems* **8**:2, pp. 117-132, March, 1997.

Querying Semistructured, Heterogeneous Information (with Dallan Quass, A
semantics. Also, a A shorter Version that appeared in DOOD '95.

Author	Publication	Year
Grishman	Information Extraction...	2006
...	...	...

9

Other motivations: price comparison



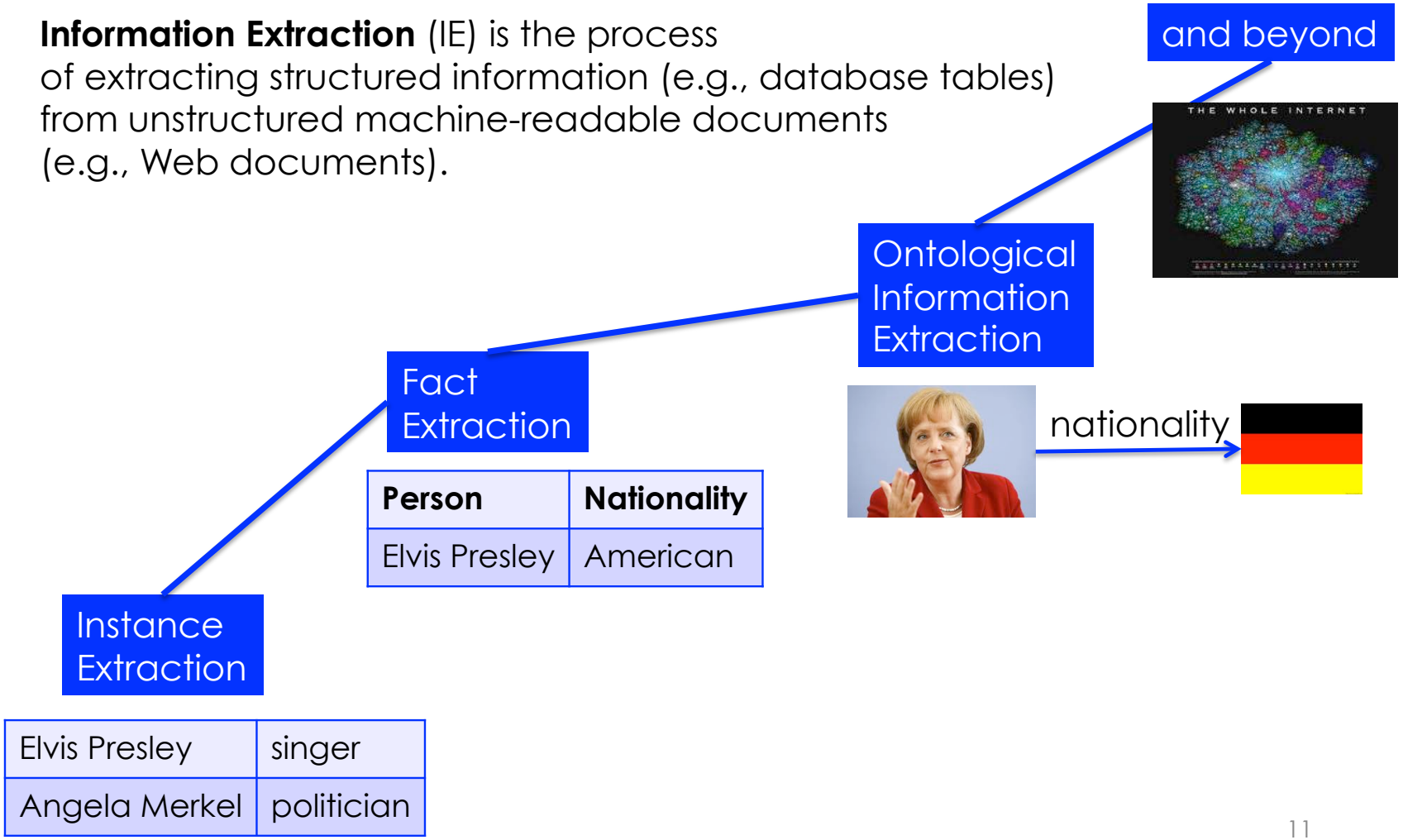
Product	Type	Price
Dynex 32"	LCD TV	$1000
...	...	

10

Table of contents

Roadmap

Information Extraction (IE) is the process of extracting structured information (e.g., database tables) from unstructured machine-readable documents (e.g., Web documents).



Instance extraction with Hearst patterns

- Entities can be extracted and categorized by automatic extraction of simple patterns:
  - Many scientists, including Einstein, believed...
  - France, Germany and other countries have been plagued with...
  - Other forms of government such as constitutional monarchy...
- Difficulties:
  - → Must parse correctly.
  - → Must be resilient to noise.

Instance extraction with set expansion

- Start with a seed set of entities of a certain type.
- Find occurrences of them at specific position in documents:
  - Lists.
  - Table columns.
- Assume that other items are other entities of the same nature.

→ Once again, this is noisy...

→ Precision and recall, see previous slides.

Set expansion example

Seed set: {Russia, USA, Australia}

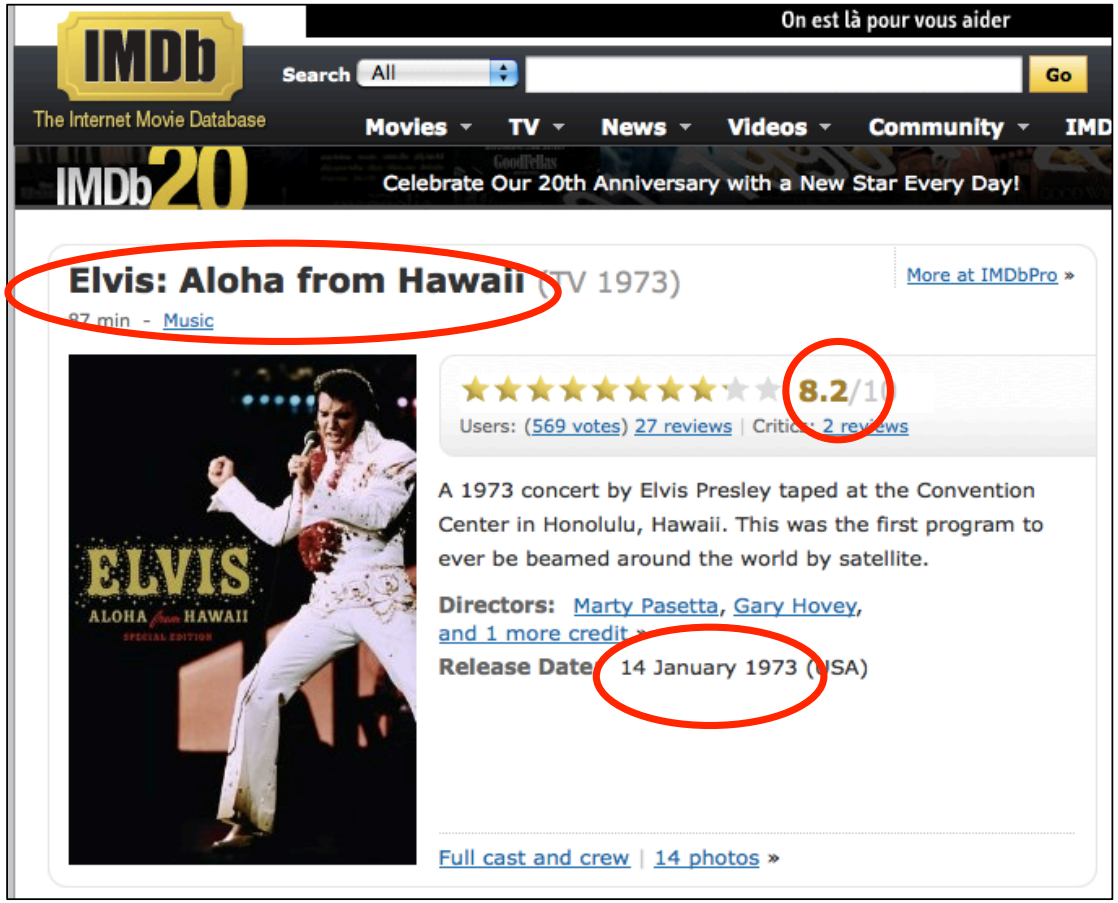
LARGEST COUNTRIES (by land mass)
locator map here
Russia 17,075,400 sq km, (6,592,846 sq mile
Canada 9,330,970 sq km, (3,602,707 sq miles
China 9,326,410 sq km, (3,600,947 sq miles)
USA 9.166,600 sq km, (3,539,242 sq miles)
Brazil 8,456,510 sq km, (3,265,075 sq miles)
Australia 7,617,930 sq km, (2,941,283 sq mil
India 2,973,190 sq km, (1,147,949 sq miles)
Argentina 2,736,690 sq km, (1,056,636 sq m
Kazakhstan 2,717,300 sq km, (1,049,150 sq
Sudan 2,376,000 sq km, (917,374 sq miles)

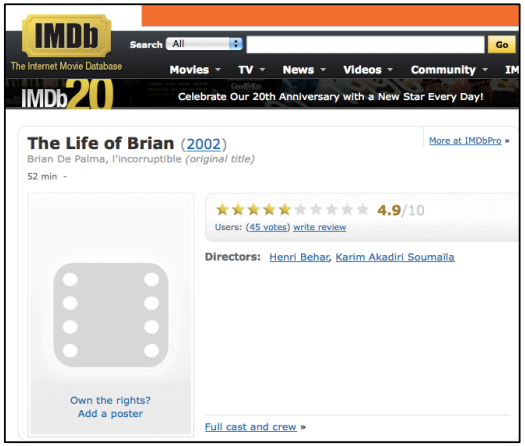
Result set: {Russia, Canada, China, USA, Brazil, Australia, India, Argentina,Kazakhstan, Sudan}

17

Fact extraction with wrapper induction

Observation: On Web pages of a certain domain, the information is often in the same spot.





28

Specifying a wrapper

- A wrapper can be expressed:
  - As a path in the DOM (usually XPath).
  - Extensions to multiple pages, e.g., OXPath.
  - As a regular expression.
- A wrapper can be produced:
  - Through manual annotation of the relevant fields.
  - Using specific knowledge of the source.
    - → Wikipedia categories and infoboxes.
  - By comparison between similar pages to find what changed.
  - Using seed pairs (known facts).
- Possibility to iterate between patterns and facts.
  - → Risk of semantic drift.

Fact extraction on text

- Entity extraction:
  - → find entities in the text.
- Named entity recognition:
  - → identify the type of entities.
    - → person
    - → organization
    - → quantity
    - → address
    - → etc.
- NLP patterns to extract facts
  - → POS patterns
  - → Parse trees.

Fact extraction on text

Einstein ha scoperto il K68, quando aveva 4 anni.

Person	Discovery
Einstein	K68

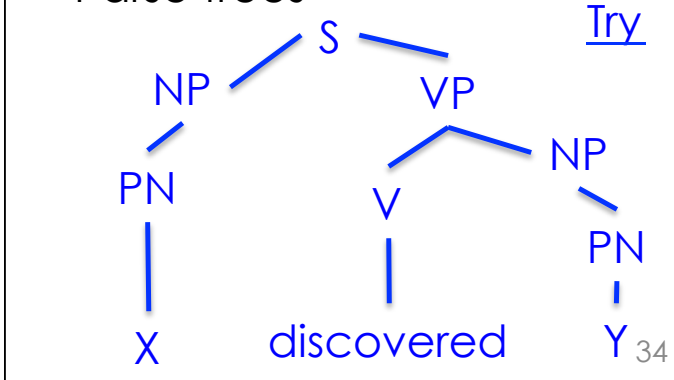
X ha scoperto il Y

Bohr ha scoperto il K69 nel anno 1960.

Person	Discovery
Bohr	K69

The patterns can be more complex, e.g.

- regular expressions
  X discovered the .{0,20} Y
- POS patterns
  X discovered the ADJ? Y
- Parse trees



34

Ontologies

- Ontology: a set of entities and relations.

Name	Ment.	Mfact	Domain	Publisher	Start	Update
YAGO	>10	>120	general	MPI	2008	2012
DBPedia[2]	4.6	3	general	OpenLink et al	2007	2014
Wikidata[3]	15	30	general	Wikimedia	2012	2014
Freebase[4]	46	2 680	general	Metaweb (Google)	2007	2014
Knowl. Vault[5]	>570?	>1 600	general	Google	2014	2014
MusicBrainz[6]	>35	>180	music	MetaBrainz	2003	2013
WordNet	0.4	2	English	Princeton	1985	2006
ConceptNet[7]	5.3	13	obvious	MIT	2000	2014

[2] http://blog.dbpedia.org/2014/09/09/dbpedia-version-2014-released/

[3] http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext#body-9

[4] http://freebase.com/

[5] http://www.newscientist.com/article/mg22329832.700-googles-factchecking-bots-build-vast-knowledge-bank.html

[6] https://musicbrainz.org/statistics

[7] http://conceptnet5.media.mit.edu/downloads/

Entity disambiguation

- Map mentions in text to entities.
- Problem: mentions are ambiguous!
  - → Use the importance of entities.
  - → Use the likelihood that a term refers to an entity.
  - → Use semantic consistency on the mappings of a document.

Entity disambiguation

- Map mentions in text to entities.
- Problem: mentions are ambiguous!
  - → Use the importance of entities.
  - → Use the likelihood that a term refers to an entity.
  - → Use semantic consistency on the mappings of a document.

(Demo: `https://gate.d5.mpi-inf.mpg.de/webaida/`)

Ontological IE

- Use the existing ontology as reference.
- Extract information from additional documents.
- Use fuzzy rules to extend the ontology (often manual):
  - → Extraction rules
  - → Logical constraints
  - → Common sense
- Reasoning:
  - → Datalog
  - → Weighted MAX-SAT
  - → Markov Logic Networks

Ontological IE

Ontology

NL documents

Elvis was born in 1935

Semantic Rules

birthdate<deathdate

First Order Logic Formulae

type(Elvis_Presley,singer)
subclassof(singer,person)
...

appears("Elvis","was born in",
"1935")
...
means("Elvis",Elvis_Presley,0.8)
means("Elvis",Elvis_Costello,0.2)
...

born(X,Y) & died(X,Z) => Y<Z
appears(A,P,B) & R(A,B)
=> expresses(P,R)
appears(A,P,B) & expresses(P,R)
=> R(A,B)
...

New facts with
1. canonic relations
2. canonic entities
3. consistency with the ontology

born → 1935

SOFIE system

Open IE

- Use the entire Web as corpus.
- Crawl the Web for new facts.
- Create new rules from what is extracted.
- Examples:
  - Open IE, University of Washington.

Open IE

- Use the entire Web as corpus.
- Crawl the Web for new facts.
- Create new rules from what is extracted.
- Examples:
  - Open IE, University of Washington.
    - → Demo: `http://openie.cs.washington.edu/`

Open IE

- Use the entire Web as corpus.
- Crawl the Web for new facts.
- Create new rules from what is extracted.
- Examples:
  - Open IE, University of Washington.
    → Demo: `http://openie.cs.washington.edu/`
  - Read the Web, CMU.
    → Demo: `http://rtw.ml.cmu.edu/rtw/`

Table of contents

Motivation

- Having structured data is nice.
- However, independent sources are not useful.
- We need to create links between data sources.
  - → Run a query across multiple relevant data stores.
  - → Perform complex transactions (booking a flight, a hotel...).
  - → Rich data visualization (integrating e.g. maps and statistics).
- We need to define the semantics of the data.
- We need to enforce constraints.
- We need to evaluate complex queries over multiple sources.

The Semantic Web

URIs	Globally unique identification of entities and relations.
OWL	Constraint language over structured data.
RDF	Storage format for structured data.
SPARQL	Query language for structured data.
LOD	Linked Open Data: draw links between data sources.

URIs

- Uniform Resource Identifier
- Like URLs.
- Not always dereferenceable.
- URNs: `urn:isbn:0486415864`
- URLs, often with namespaces:
  - → `dbp:Paris` for `http://dbpedia.org/resource/Paris`

RDF

- Resource Description Framework.
- Triples: Subject predicate object.
- `<dbp:Paris> <dbp:country> <dbp:France>`
  → The country of Paris (DBPedia resources).
- `<dbp:Paris> <foaf:homepage> <http://www.paris.fr>`
  → The homepage of Paris (FOAF relation, website).
- `<dbp:Paris> <foaf:name> "Paris"@en`
  → The name of Paris (FOAF relation, literal value).
- Multiple serializations.

RDFS

- RDF Schema.
- `<dbp:Paris> <rdf:type> <dbp:Settlement>`
  → Paris is a settlement.
- `<dbp:Settlement> <rdfs:subclassOf> <dbp:Place>`
  → If I am a `Settlement` then I am a `Place`.
- `<dbp:writer> <rdfs:subPropertyOf> <y:created>`
  → If you are the writer of something (for DBPedia)
  then you are the creator of that thing (for YAGO).

OWL

- Ontology Web Language.
- `<dbp:birthPlace> <rdf:type> <owl:FunctionalProperty>`
  → People are born in at most one place.
- `<dbp:Person> <owl:disjointWith> <dbp:Settlement>`
  → Something cannot be both a `Person` and a `Settlement`.
- `<schema:spouse> <owl:equivalentProperty> <dbp:spouse>`
  → `spouse` in Schema.org in DBPedia are equivalent properties.
- `<myonto:p4242> <owl:sameAs> <dbp:Douglas_Adams>`
  → Assert equalities between resources.

SPARQL

- SPARQL Protocol And RDF Query Language.
- Query language for RDF.

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
   ?x abc:cityname ?capital   ;
       abc:isCapitalOf  ?y  .
   ?y abc:countryname ?country ;
       abc:isInContinent  abc:Africa   .
}
```

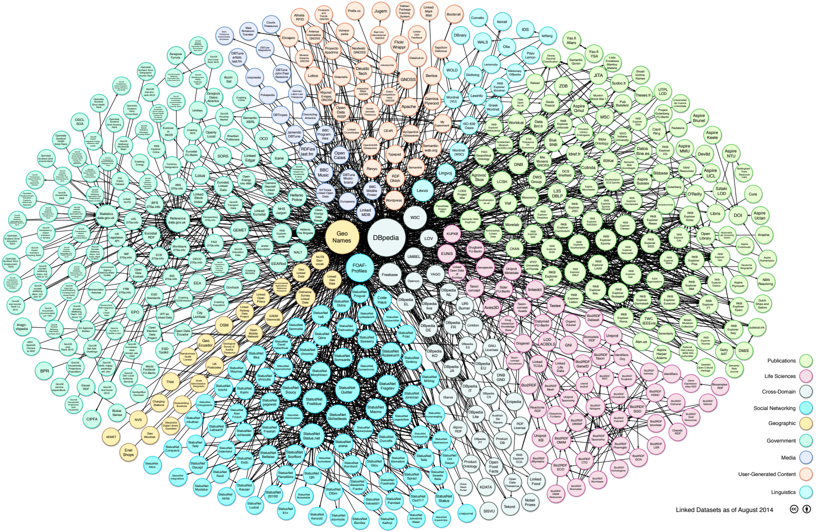
SPARQL

- SPARQL Protocol And RDF Query Language.
- Query language for RDF.

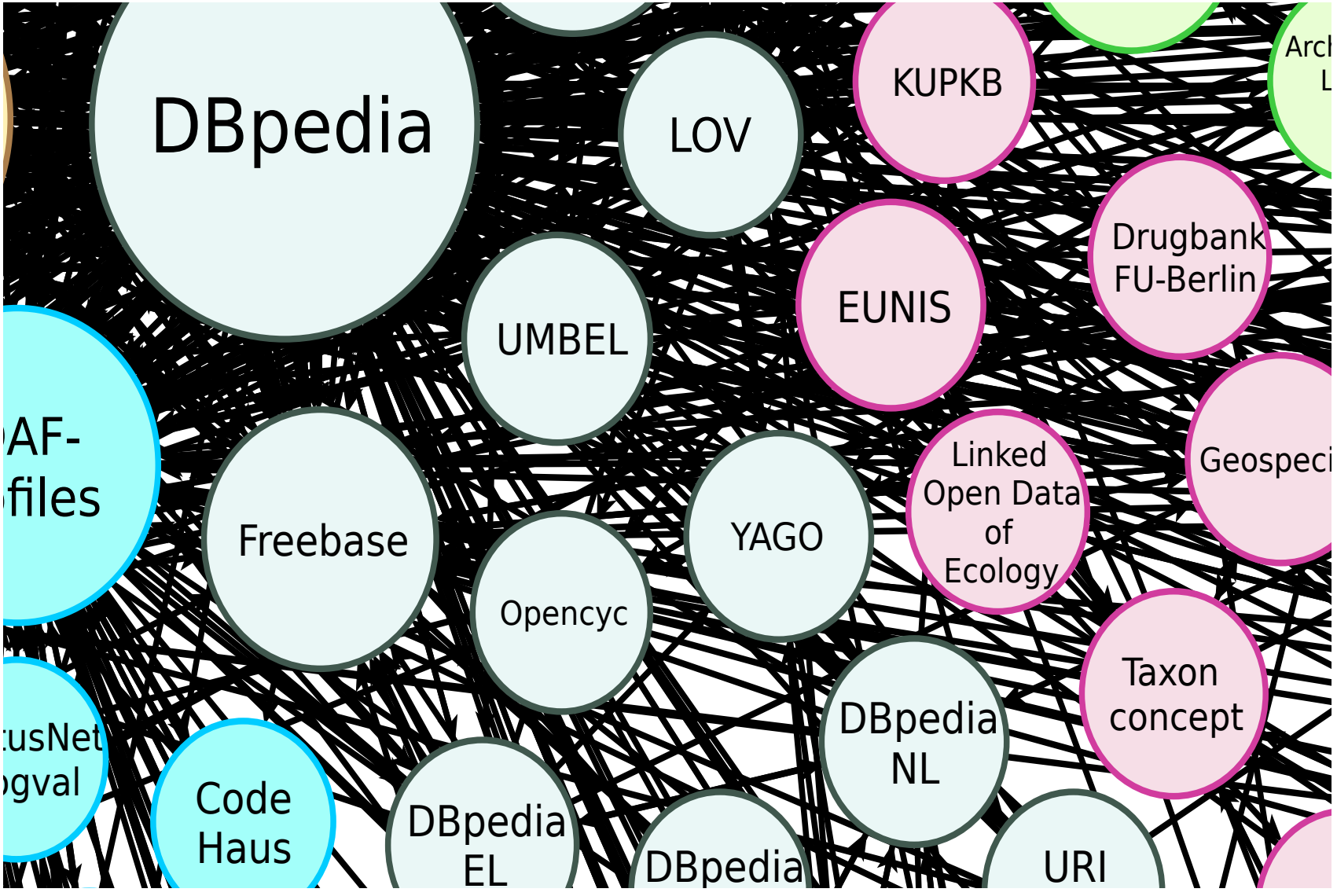
```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital   ;
     abc:isCapitalOf  ?y  .
  ?y abc:countryname ?country ;
     abc:isInContinent  abc:Africa  .
}
```

(Demo: `http://dbpedia-live.openlinksw.com/sparql/`)

Linked Open Data



Linked Open Data (zoom)



Linked Open Data

- Integrate many sources:
  - → General ontologies.
  - → Domain-specific ontologies.
  - → Open data dumps.
  - → Existing relational databases.
- Find links
  - → Automatically.
  - → By hand (relations, rules...).
- Statistics:
  - → Hundreds of ontologies.
  - → 504 million RDF links (2011).
  - → 52 billion triples (2012).[8]
  - → 5 billion entities (2012, incl., e.g., 854 million people).

→ Wealth of data, still underused

[8] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Challenges

- Manage trust and attribution.
- Manage time.
- Manage uncertainty throughout the pipeline.
  - → Extraction.
  - → Trust.
  - → Intrisic uncertainty.
- Manage complex facts.
  - → Reification.
- Manage noise.
- Compute alignments.