Information extraction

9. Applications

Simon Razniewski
Winter semester 2019/20

Announcements

- Results assignment 8
- Evaluation (w/ break in middle)
- Tentative exam schedule
  - 15 minutes/exam
  - Sample questions → today's lab
  - Conflicts? → Forum

	Exam schedule	
	14.1.	15.1.
9:00-9:20	2558462	
9:20-9:40	2576748	
9:40-10:00	2549786	
10:00-10:20	2581455	
10:20-10:40	2572706	
10:40-11:00	2571690	
11:00-11:20	2562559	
11:20-11:40	2553344	
11:40-12:00	2576861	
14:00-14:20	2564409	2581266
14:20-14:40	2558667	2576796
14:40-15:00	2576770	2576572
15:00-15:20	2571656	2581370
15:20-15:40	2548617	2565094
15:40-16:00	2576610	2550309
16:00-16:20	2579810	2571663
16:20-16:40	2576612	2550421
16:40-17:00	2561347	2572758
17:00-17:20	2568227	2570975
17:20-17:40	2576611	

Assignment 8 - Sample rules

sibling(V0, V1) :- sibling(V0, V2), sibling(V2, V1)	sup:	647
sibling(V0, V1) :- father(V0, V2), child(V2, V1)	sup:	574
sibling(V0, V1) :- sibling(V1, V0)	sup:	540
sibling(V0, V1) :- mother(V0, V2), child(V2, V1)	sup:	337
allegiance(V0, V1) :- place(V0, V2), ruler(V2, V1)	sup:	286
child(V0, V1) :- child(V0, V2), sibling(V2, V1)	sup:	278
allegiance(V0, V1) :- sibling(V0, V2), allegiance(V2, V1)	sup:	238
child(V0, V1) :- spouse(V0, V2), child(V2, V1)	sup:	232
allegiance(V0, V1) :- sibling(V2, V0), allegiance(V2, V1)	sup:	229
child(V0, V1) :- spouse(V2, V0), child(V2, V1)	sup:	228
place(V0, V1) :- allegiance(V0, V2), seat(V2, V1)	sup:	226
allegiance(V0, V1) :- child(V2, V0), allegiance(V2, V1)	sup:	208
allegiance(V0, V1) :- father(V0, V2), allegiance(V2, V1)	sup:	199
culture(V0, V1) :- sibling(V0, V2), culture(V2, V1)	sup:	186
status(V0, V1) :- sibling(V0, V2), status(V2, V1)	sup:	182
culture(V0, V1) :- sibling(V2, V0), culture(V2, V1)	sup:	175

Outline

1. Academic projects
   - Scraping and Harvesting
   - Pattern-based text extraction and OpenIE
2. Industrial Knowledge Bases
3. Knowledge Base Question Answering
4. Semantic Web

DBpedia (2007)

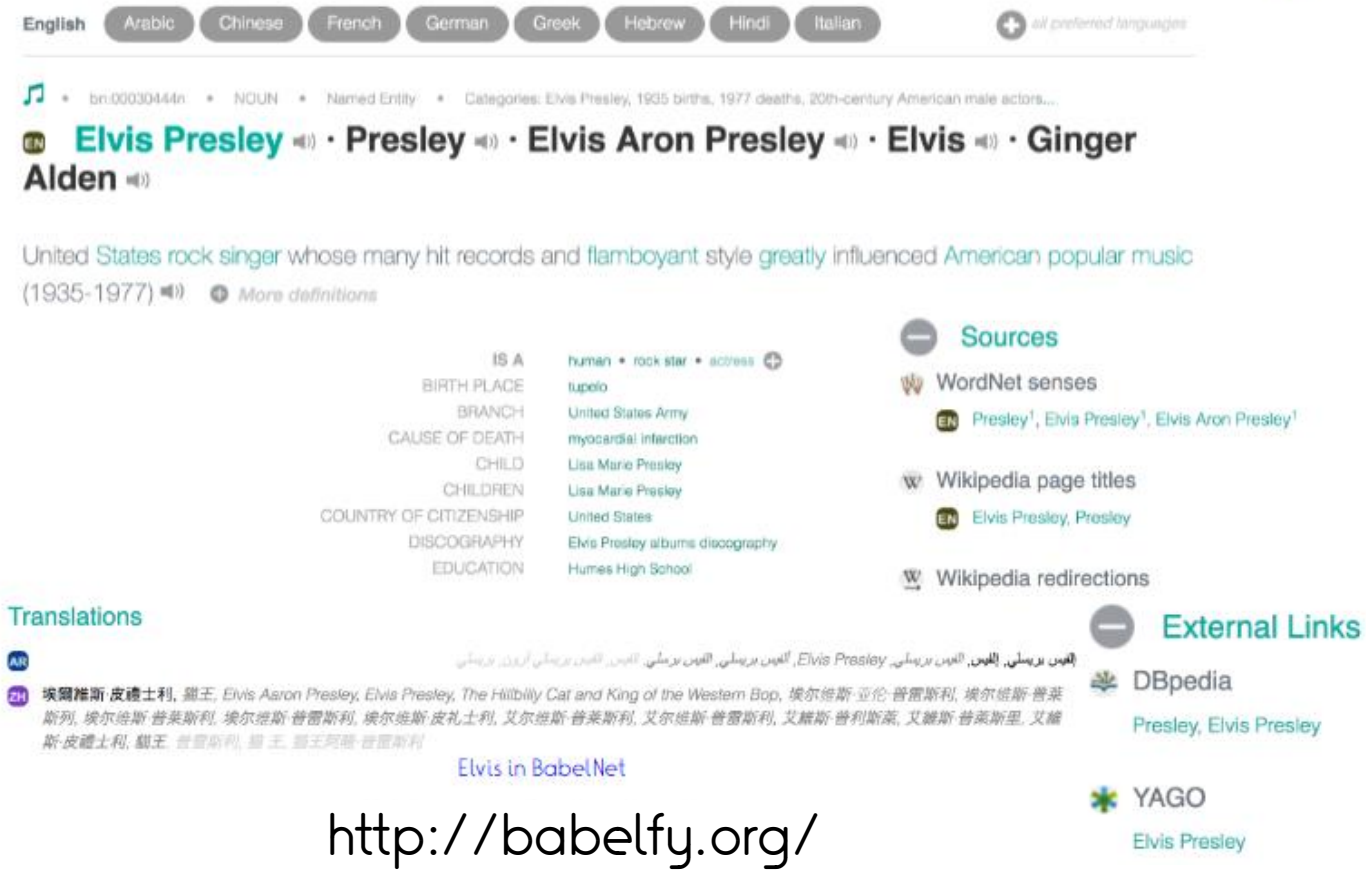
- Large-scale Wikipedia infobox+category scraping
- Manually designed mappings to consolidate synonymous attributes
- See lecture/assignment 3
- Multilingual
- No persistent IDs
- For long considered the "core" of Semantic Web (see later)
- Data access
  - Per entity: http://dbpedia.org/page/Max_Planck_Institute_for_Informatics
  - SPARQL endpoint:
    - http://dbpedia.org/snorql/?query=SELECT+%3Fitem+WHERE+%7B%0D%0A%3Fitem+dbo%3AalmaMater+dbr%3ASaarland_University%0D%0A%7D
  - Data dumps
    - https://wiki.dbpedia.org/develop/datasets
    - https://wiki.dbpedia.org/downloads-2016-10

YAGO (2007)

- Precision-oriented Wikipedia infobox+category extraction
- Subset of 76 important relations, cleaning steps (>95% precision)
- Much focus on type extraction from categories
  - "French writers" → "Writer" + "French person"
  - WordNet disambiguation and linking
- Data access
  - Per-entity access: https://gate.d5.mpi-inf.mpg.de/webyago3spotlx/Browser
  - Or https://gate.d5.mpi-inf.mpg.de/webyago3spotlxComp/SvgBrowser/
  - SPARQL access: (currently down)
  - Data dumps: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/

BabelNet (2012)

BabelNet is a multilingual lexicalized semantic network and ontology.



http://babelfy.org/

Focus on general terms, sense disambiguation, instead of named entities

Wikidata (2012)

- Largely supersedes YAGO and DBpedia
- Not itself built using automated IE techniques
  - Community generally disapproves of automated extraction
  - Isolated projects, e.g. https://github.com/google/sling
    - https://www.wikidata.org/wiki/User:Anders-sandholm
- Nonetheless highly important for IE
  - Disambiguation reference
  - Training data source (distant supervision)
- Data access:
  - SPARQL: https://w.wiki/DKU
  - Individual entities: https://www.wikidata.org/wiki/Q565400
  - JSON: https://www.wikidata.org/wiki/Special:EntityData/Q565400.json
  - Dumps: https://www.wikidata.org/wiki/Wikidata:Database_download
    - ~65 GB zipped

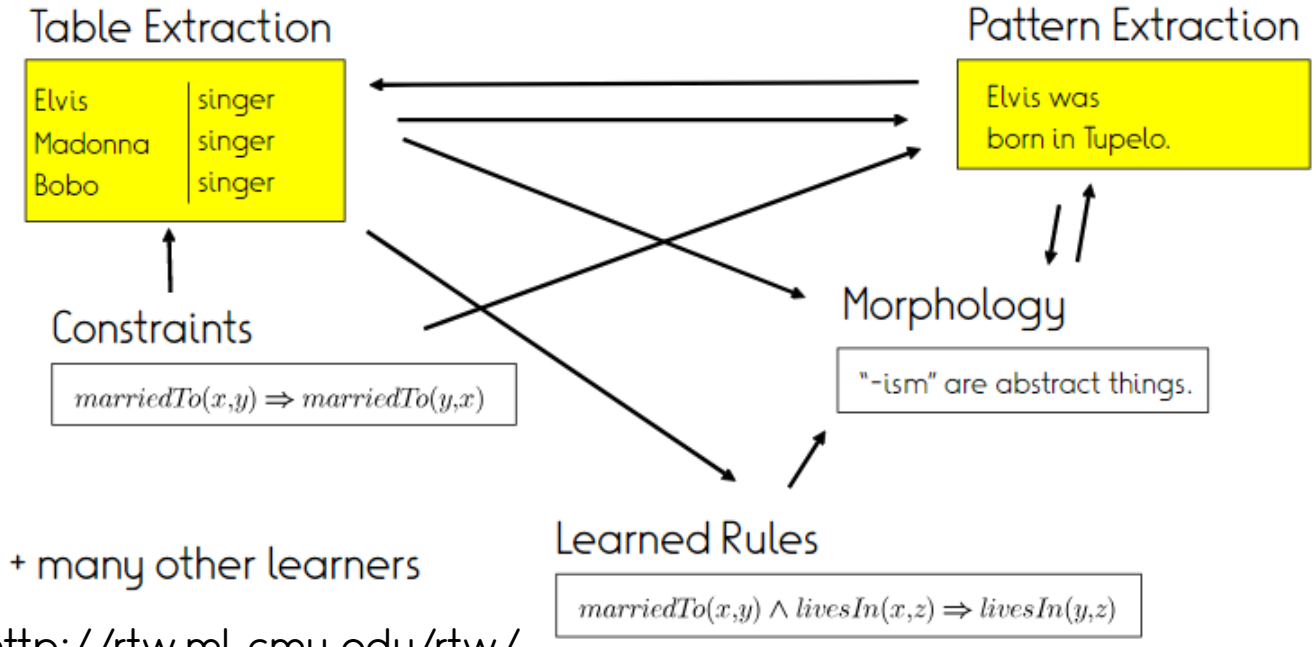
Outline

1. Academic projects
   - Scraping and Harvesting
   - Pattern-based text extraction and OpenIE
2. Industrial Knowledge Bases
3. Knowledge Base Question Answering
4. Semantic Web

NELL / Read The Web (2010)

NELL (Never Ending Language Learner) is an information extraction project at Carnegie Mellon University. It couples several learners.

327 manually designed relations each with a few curated training examples



+ many other learners

http://rtw.ml.cmu.edu/rtw/

Sales point: Continuous nature of extraction and learning

Example: NELL about "MacBook"

categories

- **product**(100.0%)
  - MBL @482 (99.9%) on 09-jan-2012 [Promotion of "product:macbook" productinstanceof "hallwayitem:windows"]
  - SEAL @7 (100.0%) on 13-jan-2010 [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74] using macbook
  - OE @806 (88.7%) on 23-jan-2014 [ ] using macbook

relations

- **createdbyagent**
  - apple001 (100.0%)
    - CPL @1024 (50.0%) on 27-oct-2016 ["arg1 iPhone and arg2"] using (apple, macbook)
- **haswikipediaurl**
  - http://en.wikipedia.org/wiki/MacBook (95.0%)
    - AliasMatcher @621 (95.0%) on 03-aug-2012 [Freebase 7/9/2012]
- **iteminvolvedwithagent**
  - apple001 (100.0%)
    - CPL @1010 (87.5%) on 04-aug-2016 ["arg1 iPhone and arg2" "arg1 releases a new version of arg2" "arg2 and iPod are trademarks of arg1"] using (apple, macbook)
- **producedby**
  - apple001 (100.0%)
    - SEAL @168 (50.0%) on 17-nov-2010 [1] using (apple, macbook)
    - OE @838 (86.3%) on 16-may-2014 [http://www.amazon.com/Apple-MacBook-MD313LL-13-3-Inch-VERSION/dp/B005CWIVYI http://macbookpro.macrumors.com/ http://macbookpro.macrumors.com/ http://macbookpro.macrumors.com/ http://macbookpro.macrumors.com/ http://macbookpro.macrumors.com/ http://macbookpro.macrumors.com/

ReVerb / OpenIE 4.0

- Knowledge base built using open information extraction
- 5 billion extractions from general web crawls
- https://openie.allenai.org/
- (previous lecture)

Open Information Extraction

Hosted by AI2 Created at

Argument 1: entity:Pyramid **Relation:**

Argument 2: All Search

165 answers from **566 sentences** (results truncated)

Pyramid

all | location (13) | fictional setting (11) | olympic participating country (9) | aircraft owner (8) | building function (6) | misc. | more types

- were built by aliens (25)
- were **Tomb** (22)
- were built by **Egyptians** (11)
- is one (11)
- is a structure (11)
- is one of **Wonders of the World** (9)
- were used as **Tomb** (9)
- is built entirely of **Limestone** (9)
- were built as **Tomb** (8)
- is in fact (8)

were built by aliens »

Extracted Synonyms:

was built by
were build by
is built by

Extracted from these sentences:

were built by

The pyramids were built by **aliens** and other scientific facts . (via ClueWeb12)
4 hours 4 hours ago Well sure , but **the pyramids** were built by **aliens** so they do n't count . (via ClueWeb12)
Which is not to say that I dismiss the possibility entirely , but it is to say that I put it in the same category with questions like , " Were **the pyramids** were built by **aliens** , " or " Will the Eagles win the NFC championship game " ? (via ClueWeb12)

Outline

1. Academic projects
   - Scraping and Harvesting
   - Pattern-based text extraction and OpenIE
2. Industrial Knowledge Bases
3. Knowledge Base Question Answering
4. Semantic Web

Industrial projects

- Google
- Microsoft
- Ebay
- Amazon
- Facebook
- IBM
- Apple
- Baidu

Google Knowledge Vault (2014)

- Ambitious project combining text extraction, semistructured extraction, and predictive models
- See lecture 8
- Usage status unknown

[Dong, Xin, et al. "Knowledge vault: A web-scale approach to probabilistic knowledge fusion, KDD 2014]

Google: Knowledge Graph (since ~2012)

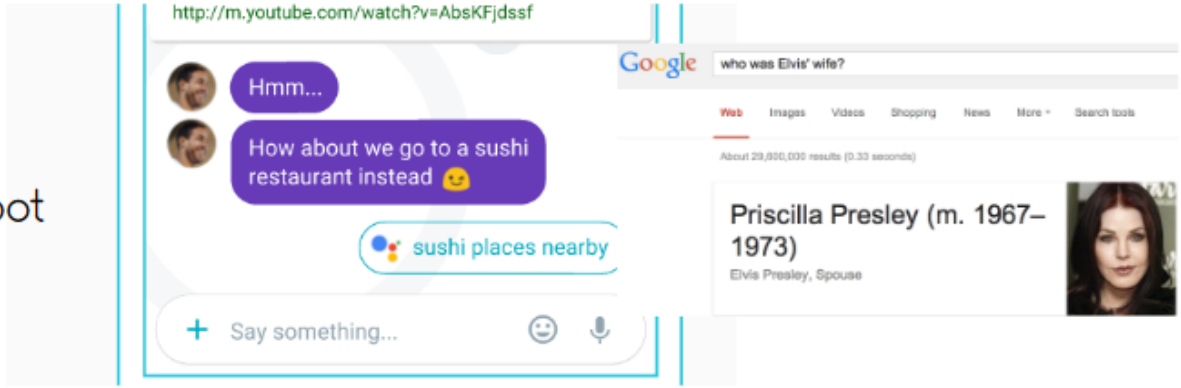
Google built its "knowledge graph", a collection of factual knowledge, from Freebase, Wikipedia, and Web sources.

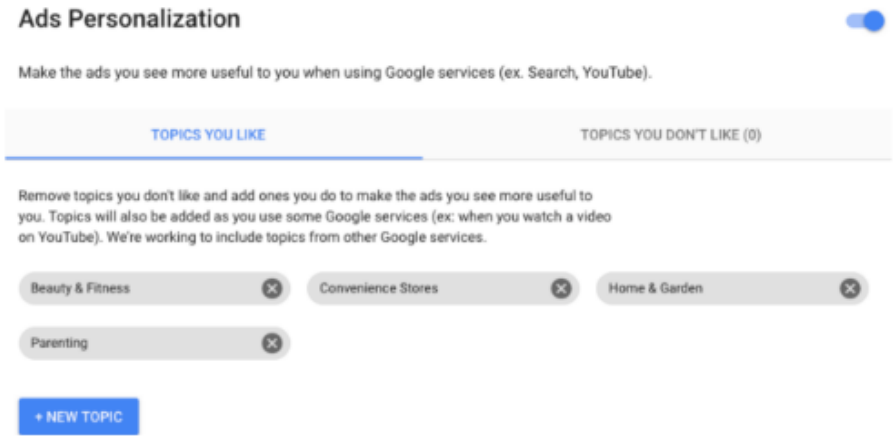
- https://developers.google.com/knowledge-graph
- Wikidata noise copied (see lecture 3)

Google: Knowledge Graph

Google uses the knowledge graph for

- Search
- Gmail
- Ads
- Its Chatbot

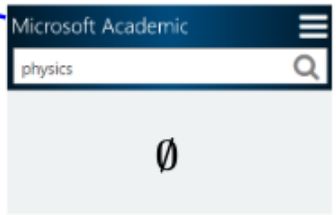




Microsoft: Satori &co

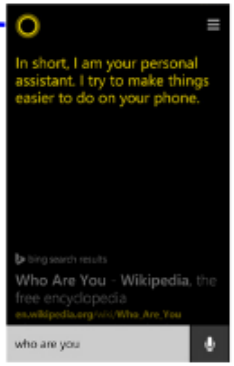
Microsoft builds

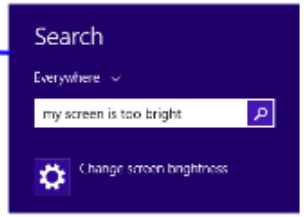
- a "world graph" (Satori)
- an academic graph
- a "work graph" based on user interactions in Office



to help

- Cortana
- search?
- Windows
- companies

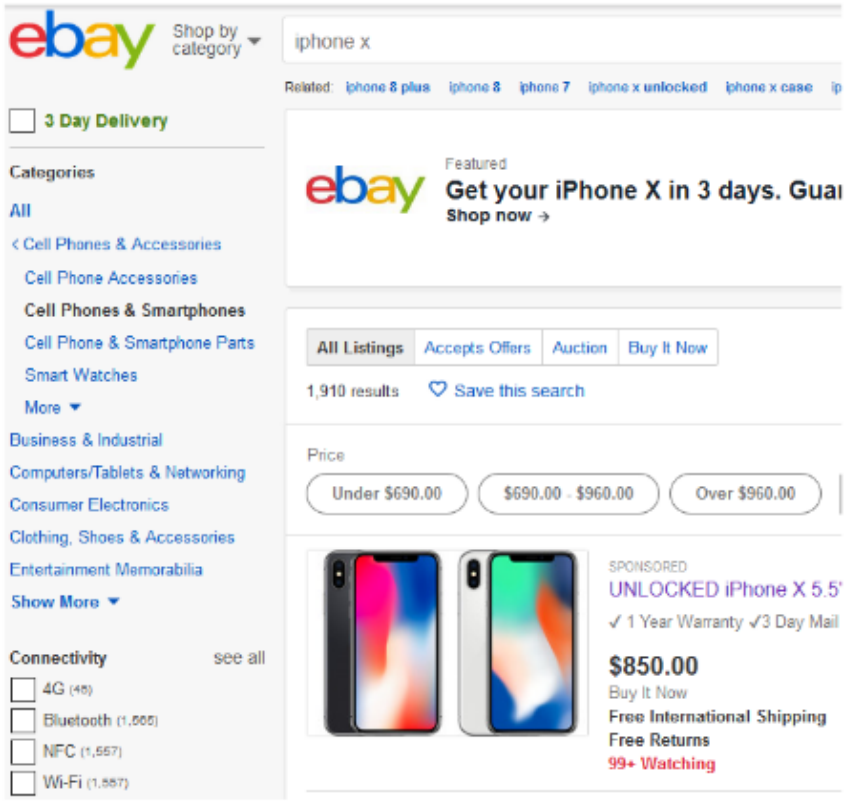




Ebay

Ebay builds a KB of
- its products
- world knowledge

in order to
- identify duplicate products
- recommend similar products



[Noy, Natasha, et al. "Industry-scale Knowledge Graphs: Lessons and Challenges." *Queue* 17.2 (2019): 20]

Amazon

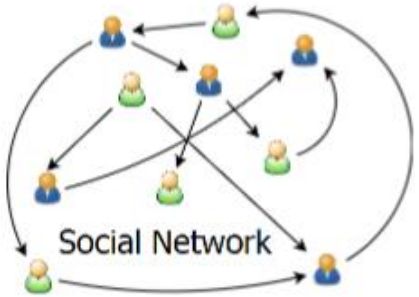
Amazon bought TrueKnowledge/Evi, a startup that built a knowledge base from Wikipedia. The knowledge base is used for Amazon Alexa/Echo.
[amazon.jobs]

amazon alexa

Facebook

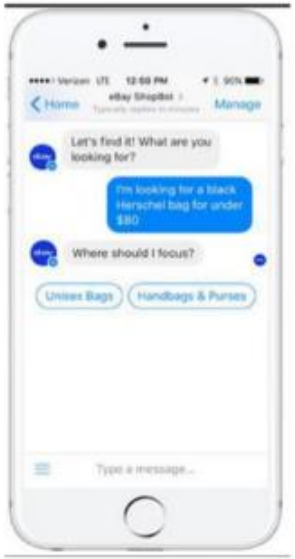
Facebook builds a KB

- of users
- of the things that users care about (celebrities, movies, etc.)



e.g., to augment messenger with

- contextual information/links
- contextual smileys
- proposed replies
- proposed actions (book taxi)



IBM: Watson

IBM sells software to build a KB to

- banks
- IT services/customer services
- defense organizations

Its showcase product is Watson.

Watson

New York Times article

Watson outperformed the 74-fold human winner in the Jeopardy quiz show

Apple?

Apple appears to use a knowledge base for Siri.

Baidu

- Non-English languages traditionally underrepresented
- Open (academic) solutions:
  - Zhishi.me: Chinese-language equivalent of DBpedia
    - Based on Baidu Baike, Hudong Baike, Chinese Wikipedia
  - Xlore: English-Chinese alignment KB
- Baidu has apparently three internal knowledge graphs
  - https://www.mdpi.com/2071-1050/10/9/3245/htm
- Huawei building a knowledge graph?
  - https://www.huawei.com/en/press-events/news/2019/9/atlas-series-products-cloud-services-all-scenario-ai-solutions

Outline

1. Academic projects
   - Scraping and Harvesting
   - Pattern-based text extraction and OpenIE
2. Industrial Knowledge Bases
3. Knowledge Base Question Answering
4. Semantic Web

Evaluation

Advertisement: Thesis topics

- http://simonrazniewski.com/#theses
  1. Social commonsense knowledge extraction
     - "Americans like guns, Germans speed on highways, Japanese bow for greetings"
     - Treasure of world knowledge, yet risk of bias and prejudice
     - Context: IE, ML
  2. Commonsense extraction from children (audio)books
     - Does infant content make commonsense more explicit?
     - Context: IE
  3. Stability and completeness prediction in Wikidata
     - What information is complete, and which one is stable?
     - Context: Data management, Machine Learning
  4. Topical image representativeness and coverage
     - What's missing in an image collection?
     - Context: Data management, (Computer Vision)
  5. Knowledge-grounded story generation
     - Can structured knowledge yield better stories?
     - Context: Text generation

- Limited availability

  → For planning please write me till Feb 17
  (actual start flexible)

Outline

1. Scraping and Harvesting
   - DBpedia, Yago, BabelNet, (Wikidata)
2. Pattern-based text extraction and OpenIE
   - NELL and ReVerb
3. Industrial Knowledge Bases
4. Knowledge Base Question Answering
5. Semantic Web

Question answering: Vital for information access

What are films directed by Nolan?

- ★ Direct answers to questions
- ★ Saves time and effort
- ★ Natural in voice UI

Christopher Nolan / Films directed

The Dark Knight
2008

Interstellar
2014

Question answering: Vital for information access

What are the Oscar nominations of Nolan?

Christopher Nolan Academy Awards Awards / Awards

Best Picture
2018 · Dunkirk

Best Director
2018 · Dunkirk

Best Picture
2011 · Inception

Best Original Screenplay
2011 · Inception

Approaches to question answering

- Traditional IR-style approach: Match question with text phrases in documents
  - "What is the capital of Belgium"
  - "Brussels is the capital of Belgium"
  - Works only for simple questions
  - Misses additional conditions
- Google, Siri, Echo et al.
  - Precision much more important than recall
  - Answer origin needs to be debuggable/explainable

  → Question answering from structured sources much preferred

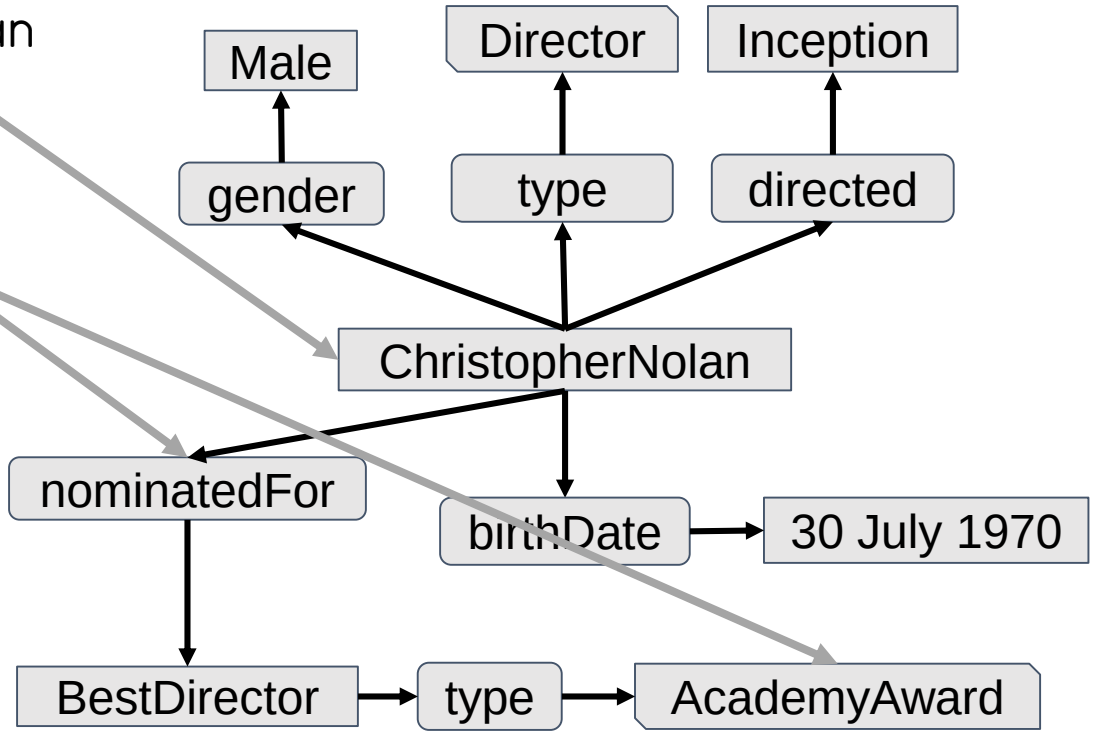
Question answering is a hot topic

- ★ QA over knowledge graphs [Abujabal et al. 2018]
- ★ Reading comprehension QA [Reddy et al. 2018]
- ★ Visual and multimodal QA [Lu et al. 2016]
- ★ Community QA [Hoogeveen et al. 2018]
- ★ Passage retrieval and sentence selection [Shen et al. 2018]
- ★ Non-factoid: Causal, procedural, ...

QA over Knowledge Graphs

Which Oscar nominations did Nolan receive?

<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>



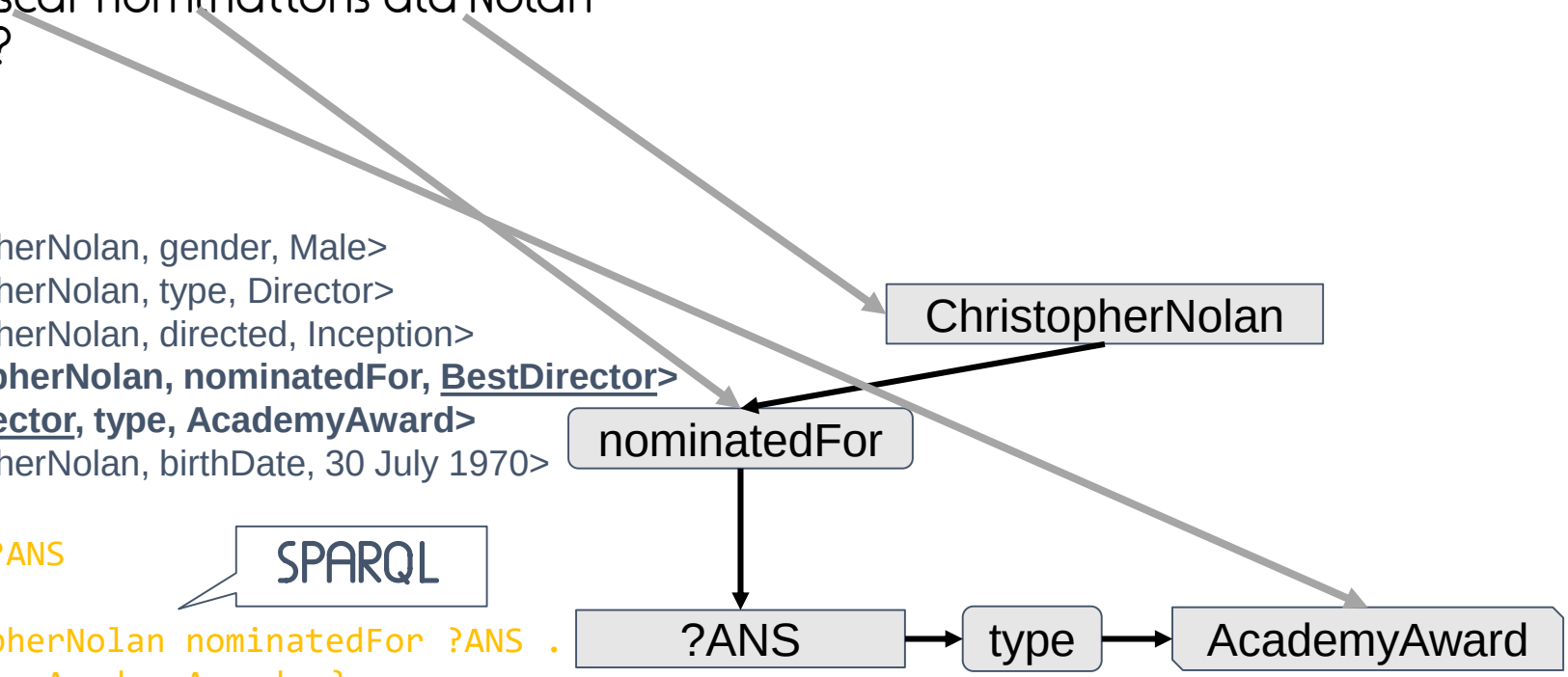
QA over Knowledge Graphs

Which Oscar nominations did Nolan receive?

<ChristopherNolan, gender, Male>
<ChristopherNolan, type, Director>
<ChristopherNolan, directed, Inception>
<ChristopherNolan, nominatedFor, BestDirector>
<BestDirector, type, AcademyAward>
<ChristopherNolan, birthDate, 30 July 1970>

```
SELECT ?ANS
WHERE {
ChristopherNolan nominatedFor ?ANS .
?ANS type AcademyAward  }
```

SPARQL



BestDirector

Generalizing QA

- If we can answer:
  - What are the Oscar award nominations of Nolan?
- Then we should be able to answer:
  - What are the Cannes award nominations of Ryan Coogler?
  - Which Oscar award nominations did Nolan receive?

Same syntax!

Same semantics!

Template-based Question Answering

★ Interpretable

Question	**Question template**
Who is **Inception**'s **director**?	Who is **<NOUN1>**'s **<NOUN2>**?
Query	**Query template**
?ANS **director Inception**	?ANS **<PRED1> <ENT1>**

1 SPARQL triple pattern

Template-based Question Answering

★ Generalizes to new domains

Who is **Libya**'s **president**?

Who is **Messi**'s **manager**?

Question	**Question template**
Who is **Inception**'s **director**?	Who is **<NOUN1>**'s **<NOUN2>**?
Query	**Query template**
?ANS **director Inception**	?ANS **<PRED1> <ENT1>**

1 SPARQL triple pattern

Template-based Question Answering

★ Generalizes to new domains

Question	**Question template**
Who **plays** the **role** of **Cobb** **in** **Inception**?	Who **<VERB>** <DT> **<NOUN>** **<PREP>** **<NOUN>**?
Query	**Query template**
?ANS **playsIn** **Inception** ?ANS **role** **Cobb**	?ANS **<PRED1>** **<ENT1>** ?ANS **<PRED2>** **<ENT2>**

2 SPARQL triple patterns

Challenges with templates

- ★ Hand-crafted by experts

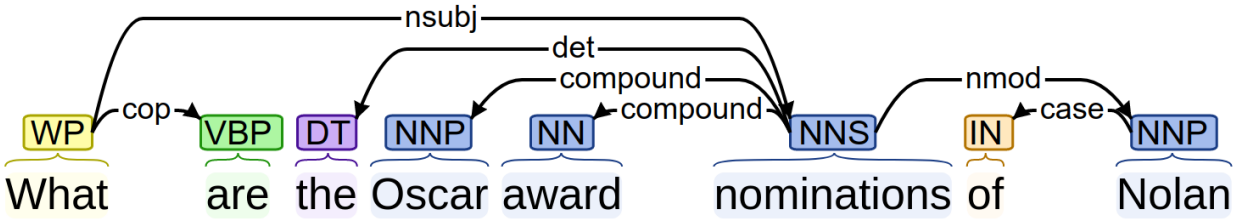
  (Fader et al. 2014; Unger et al. 2013)

  - ★ Low coverage
- ★ Solution: Learn templates
  - ○ Question templates
  - ○ Query templates
  - ○ Slot alignments

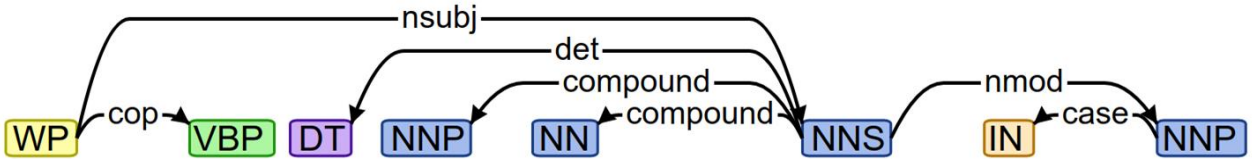
Dependency-parse-based templates

Question: What are the Oscar award nominations of Nolan?

Dependency parse

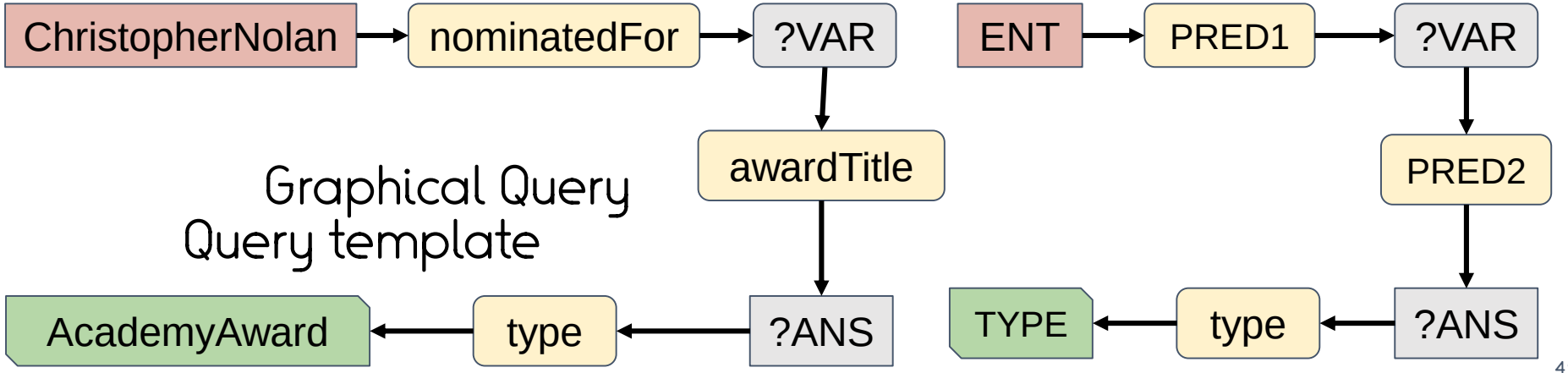


Question template (labeled nodes and edges)

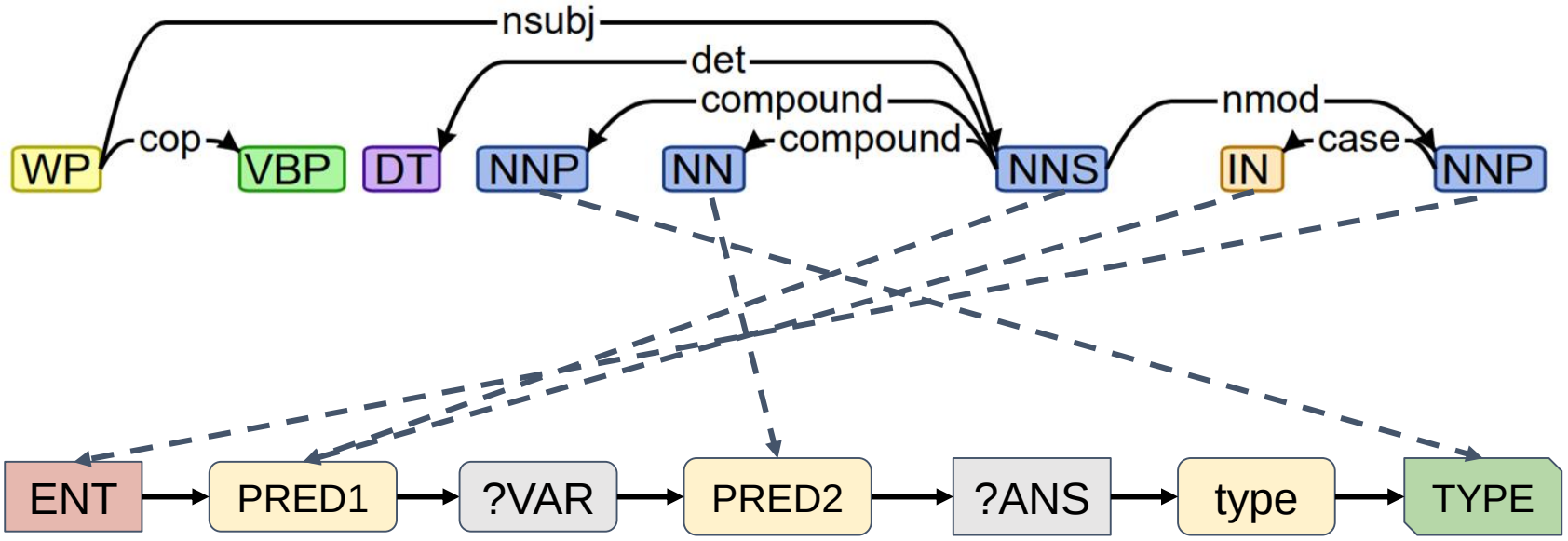


Graphical query templates

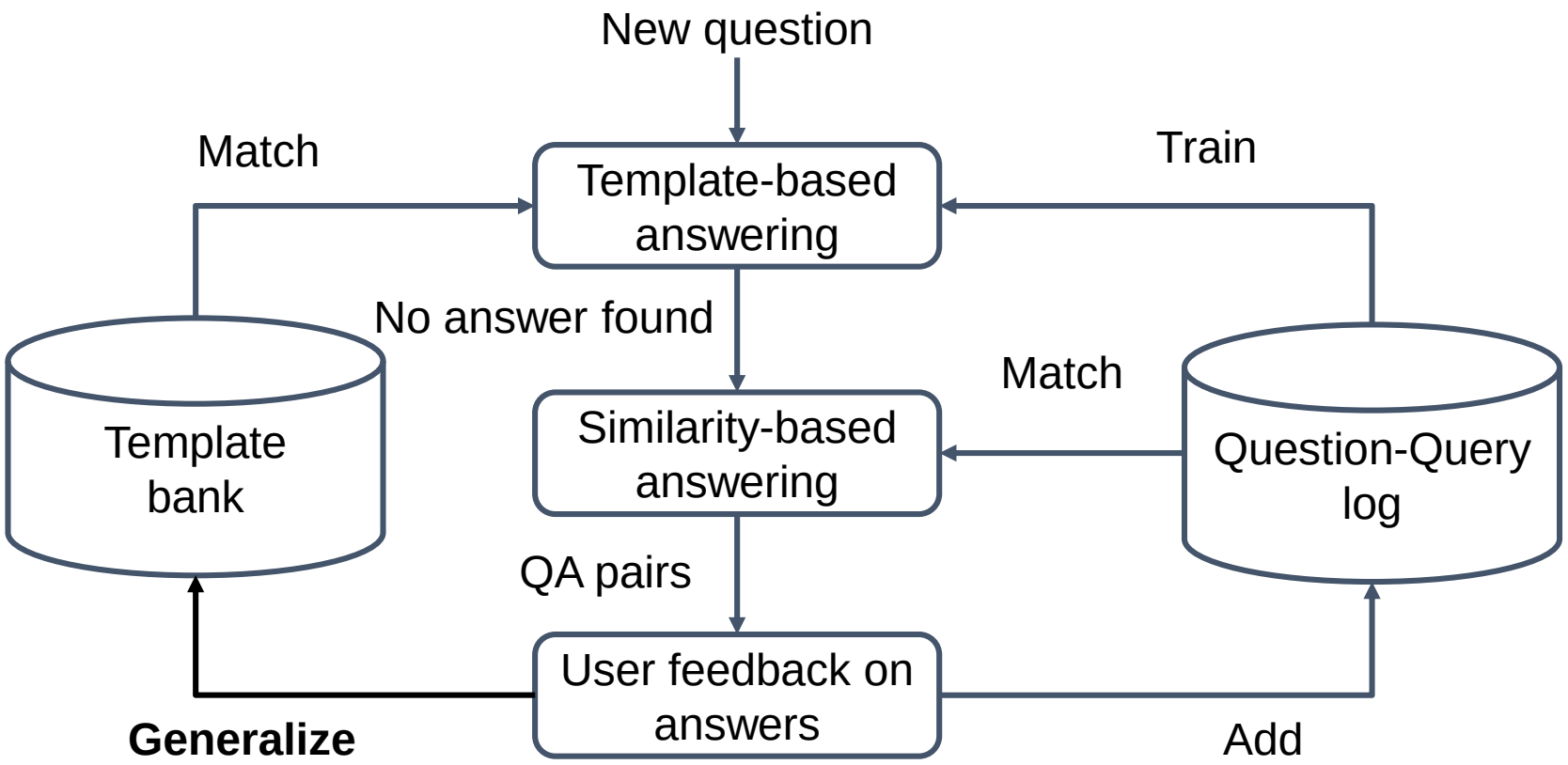
Query: ChristopherNolan nominatedFor ?VAR .
?VAR awardTitle ?ANS .
?ANS type AcademyAward



Slot Alignments



Template-based question answering



Training template-based QA

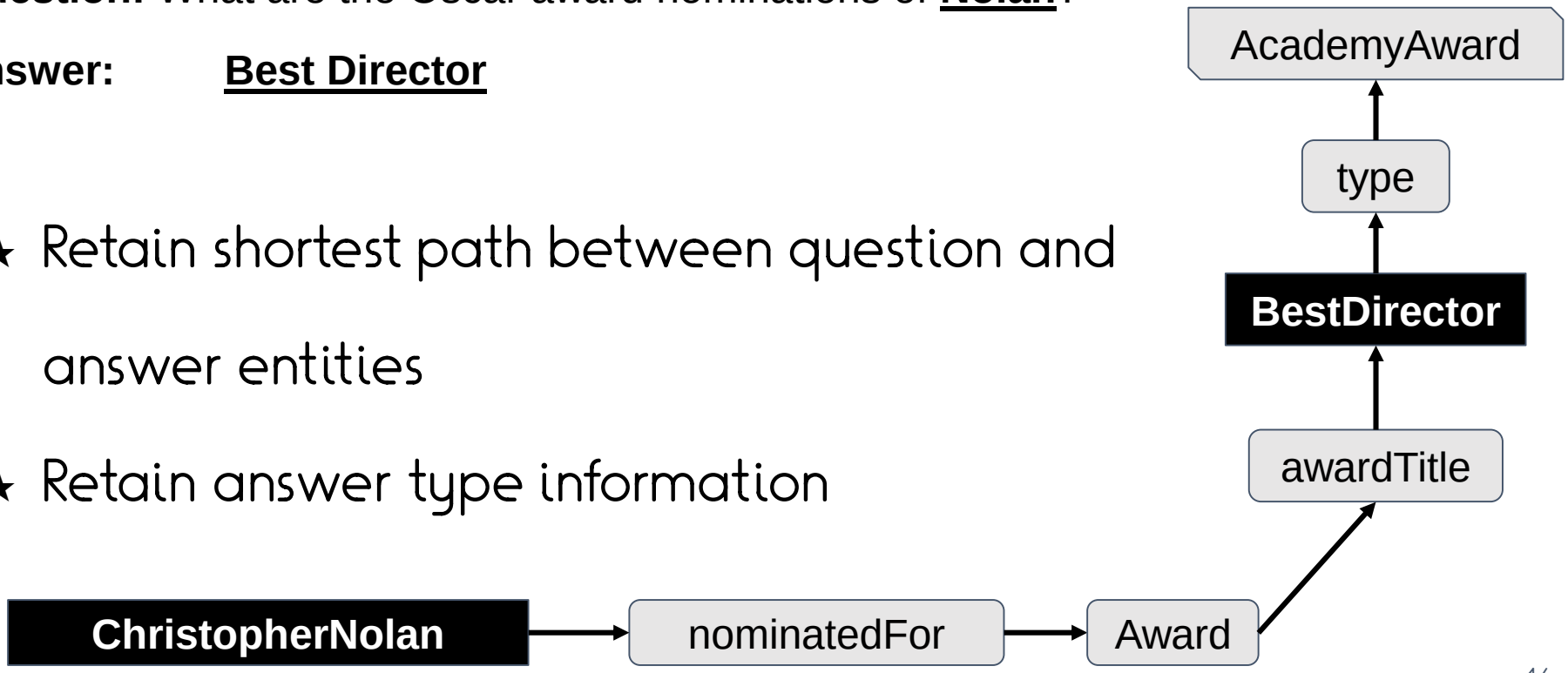
- ★ Collecting question-query pairs difficult
- ★ Start with question-answer pairs instead
- ★ Create queries by distant supervision
- ★ Generalize to create slot-aligned templates

Distant supervision from Q-A pairs

Question: What are the Oscar award nominations of **Nolan**?

Answer: **Best Director**

- ★ Retain shortest path between question and answer entities
- ★ Retain answer type information



Distant supervision from Q-A pairs

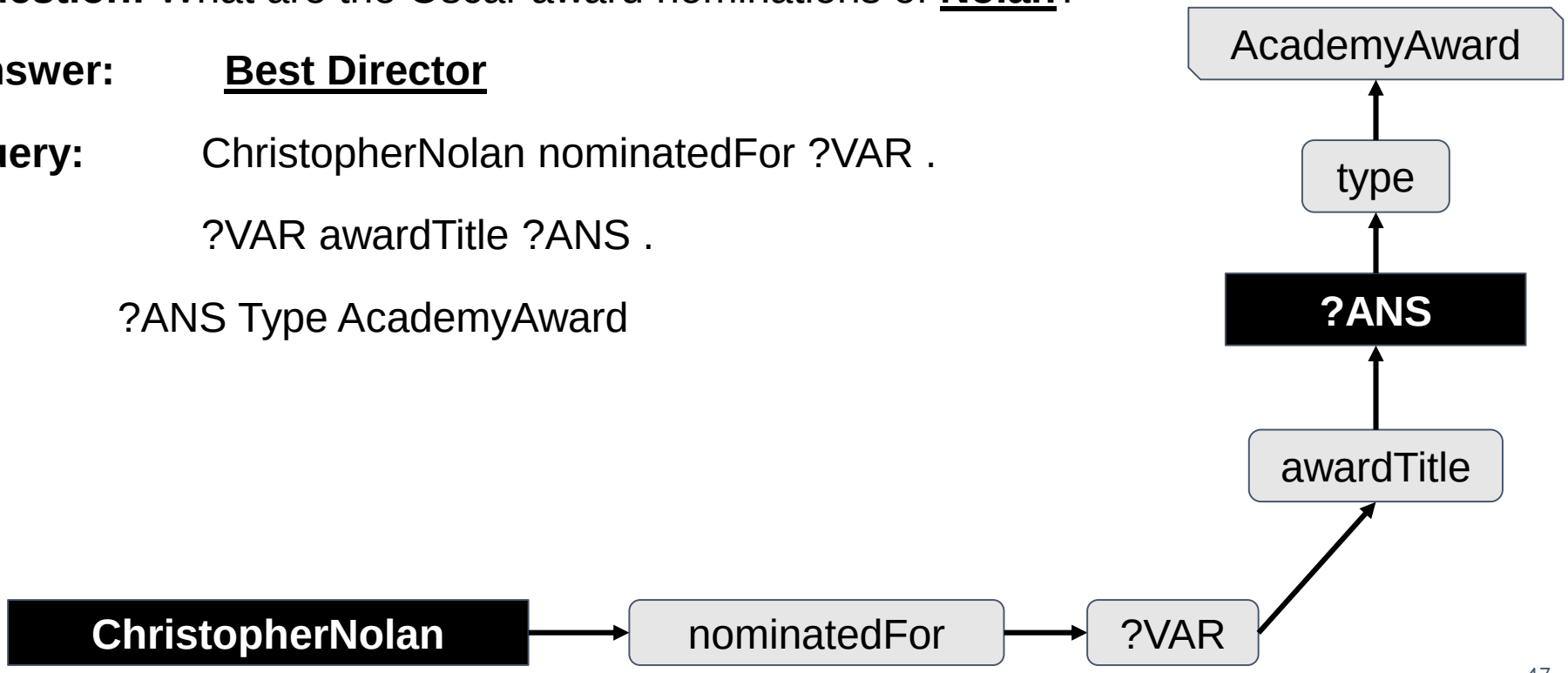
Question: What are the Oscar award nominations of **Nolan**?

Answer: **Best Director**

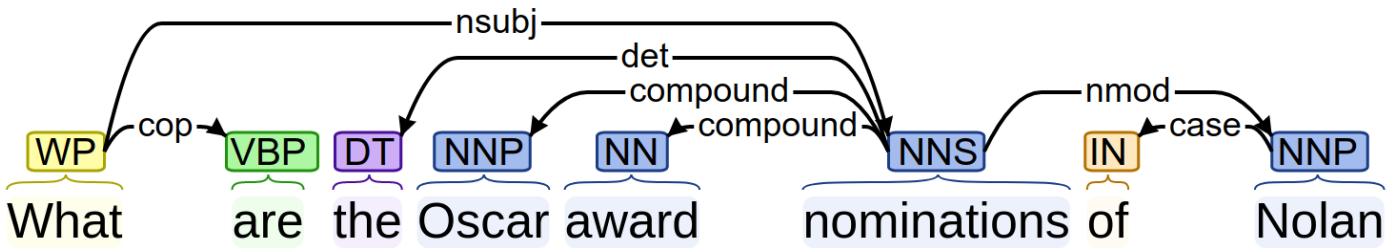
Query: ChristopherNolan nominatedFor ?VAR .

?VAR awardTitle ?ANS .

?ANS Type AcademyAward



Question-schema alignment



Question

what nominations | oscar nominations | oscar award | nominations of | oscar

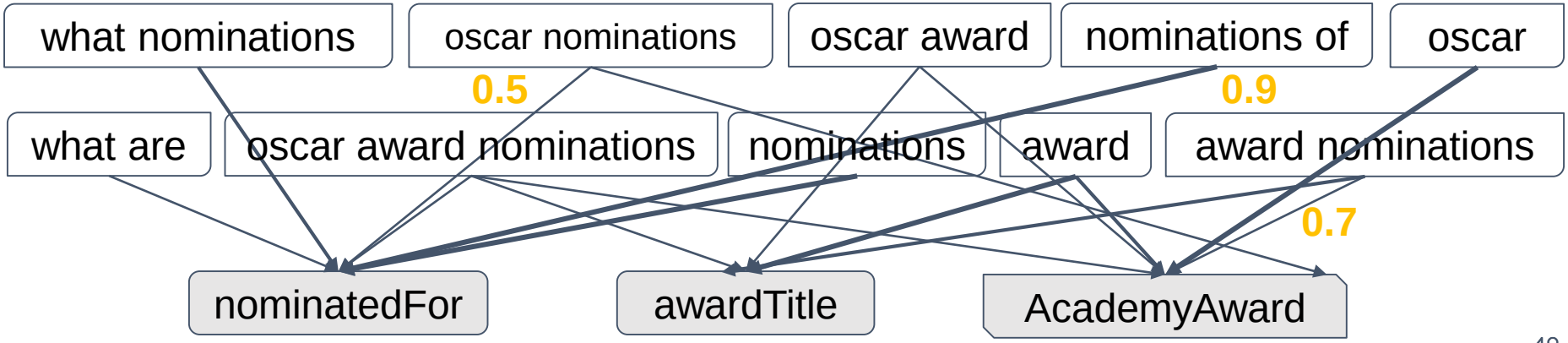
what are | oscar award nominations | nominations | award | award nominations

KB schema

nominatedFor | awardTitle | AcademyAward

Create Candidate Alignments

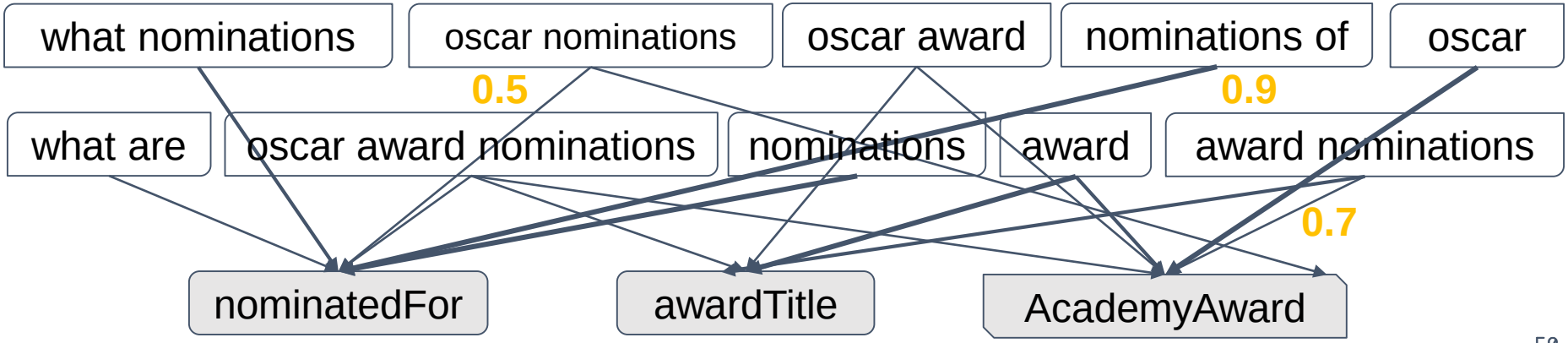
- **Bipartite graph** with edge weights (Yahya et al. 2012)
- **Weights** from lexicons L_P and L_T (Abujabal et al. 2017, Berant and Liang 2013)



Create Candidate Alignments

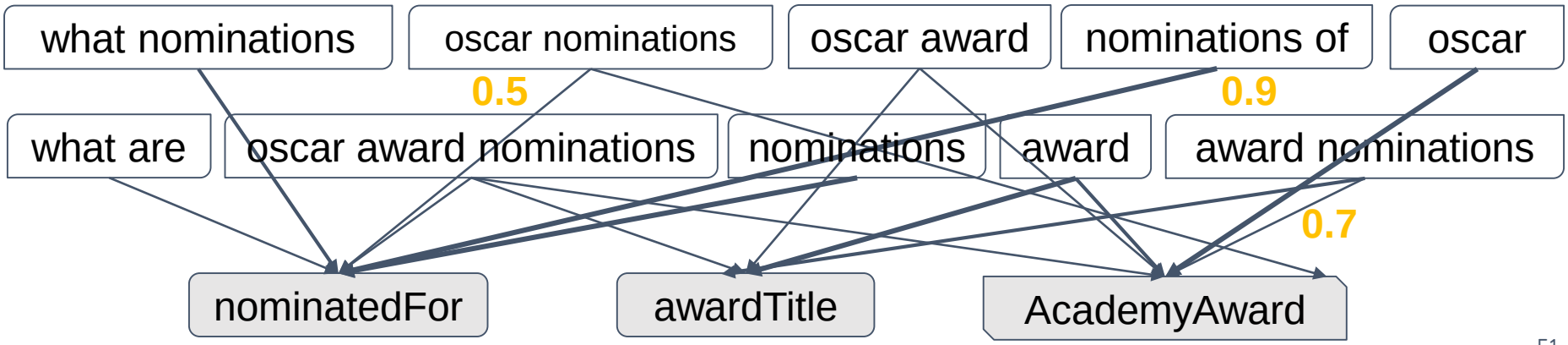
Phrase	KG Predicate	Weight
nominee for	nominatedFor	0.8
nominations of	nominatedFor	0.9
oscar nominations	nominatedFor	0.5

Phrase	KG Type	Weight
Academy Award	AcademyAward	0.9
Oscar	AcademyAward	0.7
Oscar Award	AcademyAward	0.8



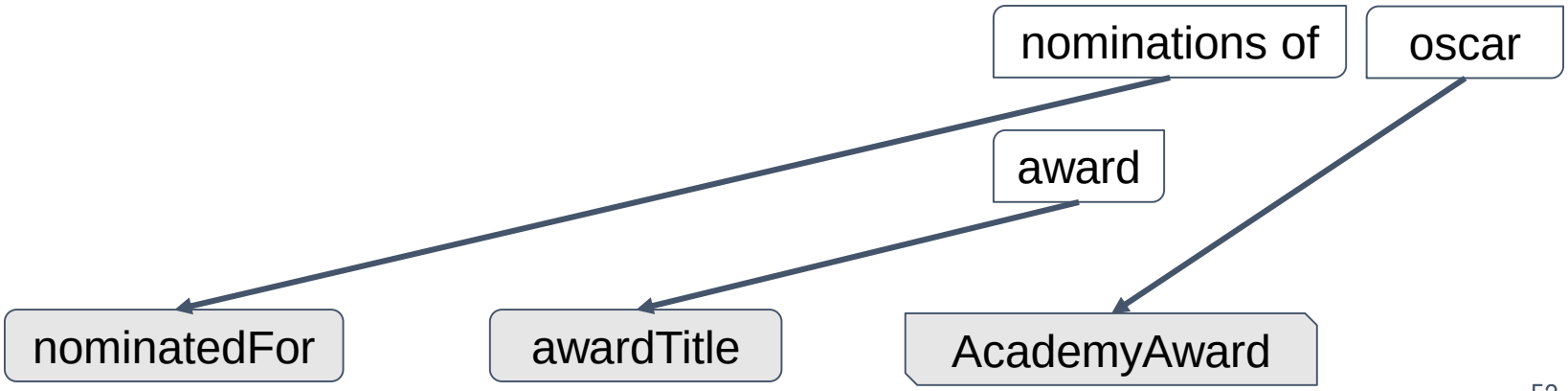
Optimal Mapping via ILP

- ★ Best alignment of items with Integer Linear Program (ILP)
  - ★ At least/at most constraints
  - ★ Type coherence

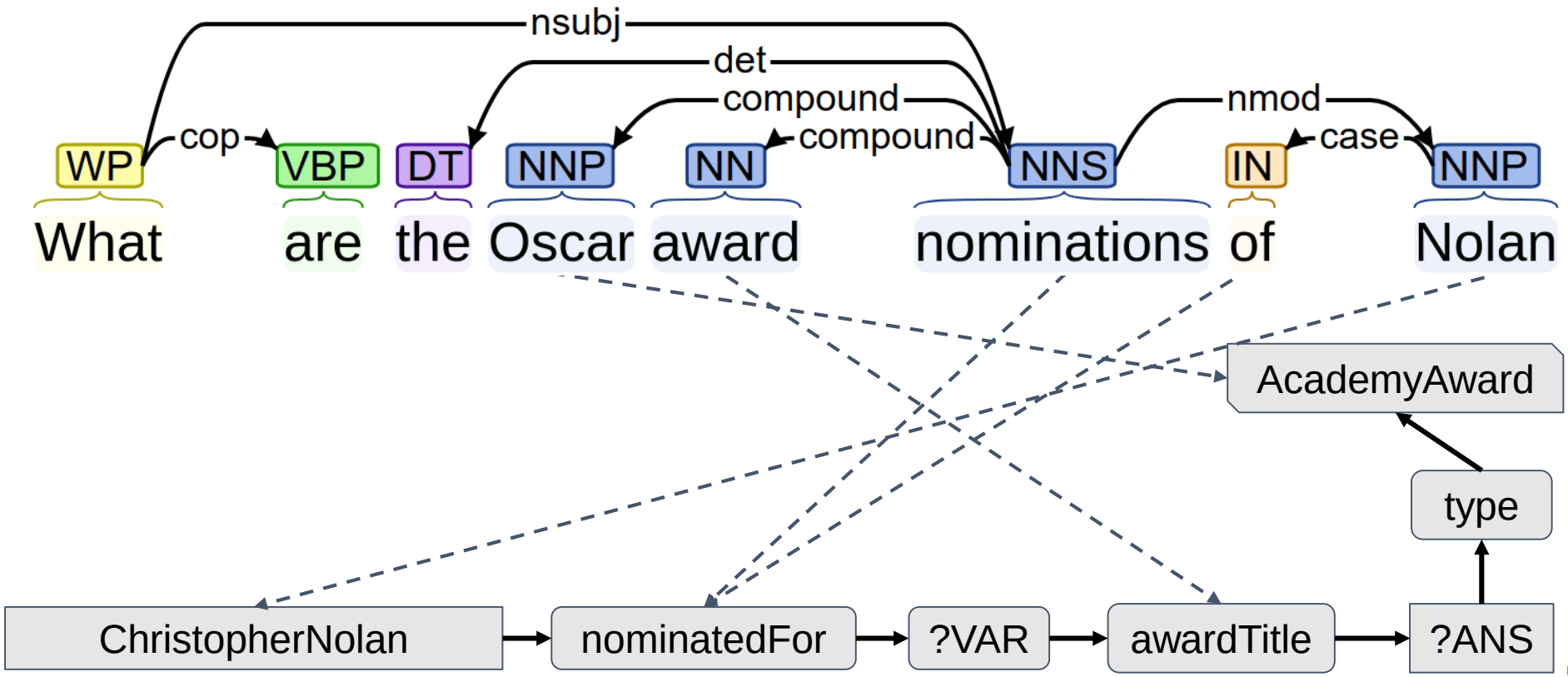


Optimal Mapping via ILP

- ★ Best alignment of items with Integer Linear Program (ILP)
  - ★ At least/at most constraints
  - ★ Type coherence

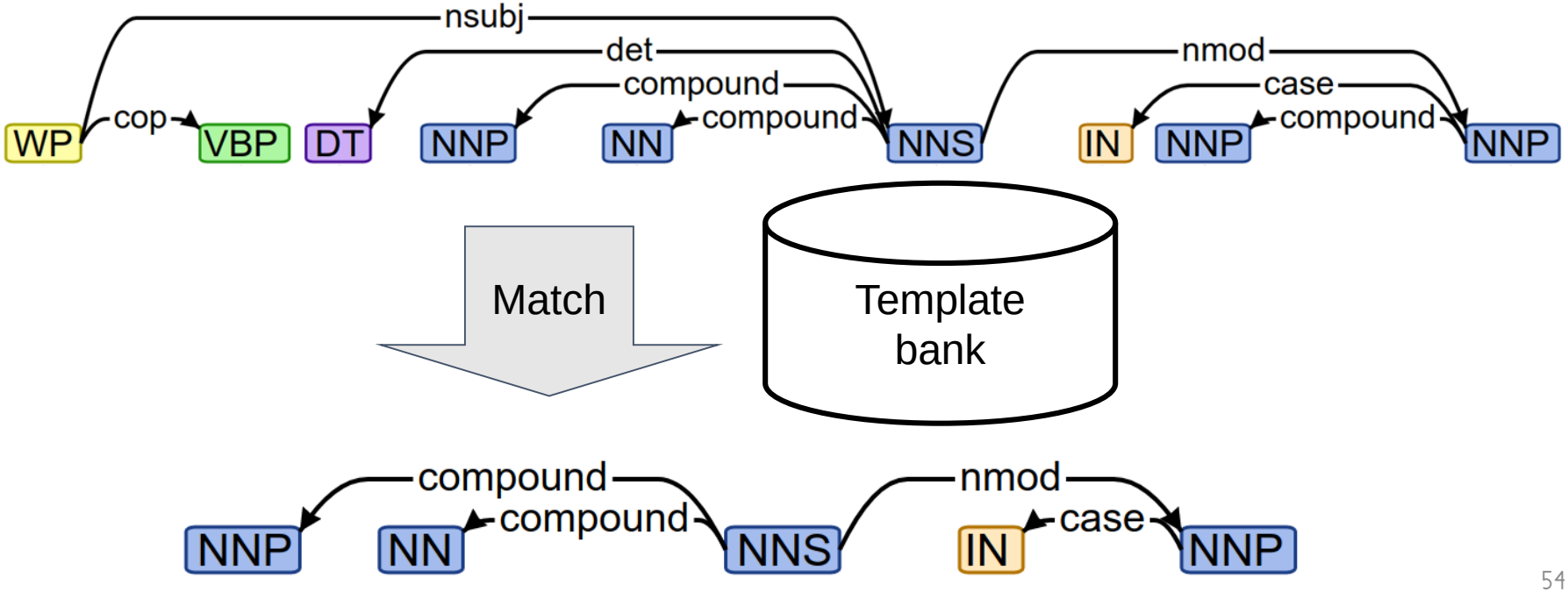


Apply Alignment to Question-Query

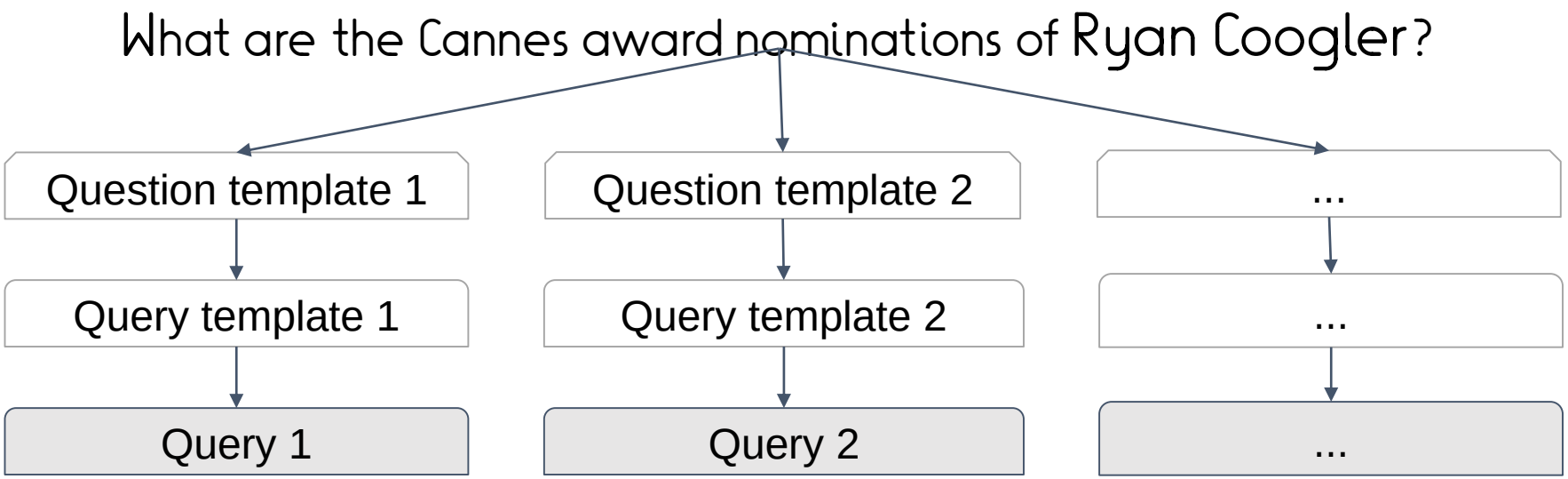


Answering with templates

New question: What are the Cannes award nominations of Ryan



Instantiating Queries



RyanCoogler nominated
?VAR .
?VAR awardTitle ?ANS .
?ANS Type CannesAward

RyanCoogler awarded
?VAR .
?VAR awardTitle ?ANS .
?ANS Type GoldenGlobe

Rank queries with learning to rank and execute best query

Closing the Loop with User Feedback

- ★ So far, assumed all answers were correct: Pseudo-relevance
- ★ Pseudo-relevance degrades quality
- ★ Users provide feedback on answers

  Question: Which Oscar nominations did Nolan receive?

  Answer: Best Director

  User: 

- ★ Positive feedback:
  - ○ Learn new template from question-query
  - ○ Add new question-query to log
  - ○ Update learning-to-rank model

Outline

1. Academic projects
   - Scraping and Harvesting
   - Pattern-based text extraction and OpenIE
2. Industrial Knowledge Bases
3. Knowledge Base Question Answering
4. Semantic Web

We can do I.E. – what now?

Airport	Location
Heathrow	London

Sources of incompatility

Airport	Location
Heathrow	London

Microsoft

?

Airport Name	City
Heathrow Airport	Londres

?

<airport>
<placeOrCity>

[Images form Wikicommons, except Oracle. Company logos for illustration only]

Where do we need interaction?

- Booking a flight
  Interaction between office computer, flight company, travel agency, shuttle services, hotel, my calendar

- Finding a restaurant
  Interaction between mobile device, map service, recommendation service, restaurant reservation

- Intelligent home
  Fridge knows my calendar, orders food if I am planning a dinner

Where do we need interaction?

- Web service composition
  Interaction between client and Web services
  and Web services themselves

- Personal assistant
  Connects calendar, email, restaurants, secretary, etc.

- Merging data after company mergers
  (e.g. Apple buys Microsoft)
  Different terminology has to be bridged,
  accounts to be merged

- Merging data in research
  e.g. biochemical, genetic , pharmaceutical research data

Def: Semantic Web

Idea: We need an infrastructure that allows computers to "understand" their data.

This infrastructure shall

- allow machines to process data from others
- ensure interoperability between schemas, devices and organizations
- allow data to describe data
- allow machines to reason on the data
- allow machines to answer semantic queries

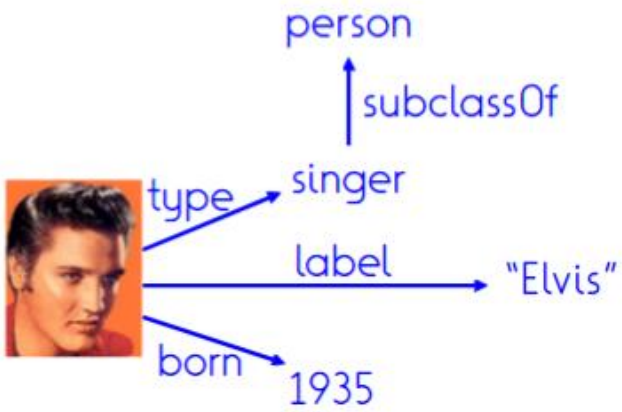
This is what the Semantic Web aims at

The Semantic Web is an evolving extension of the World Wide Web, in which data is made available in one standardized semantic format.

Reminder: RDF

RDF (Resource Description Framework) is a knowledge representation based on

- entities
- classes
- binary relations
- labels



Globally identifying entities

KB1

Elvis

KB2

Elvis

KB3

Elvis

KB4

Elvis

Def: URI

A URI (Uniform Resource Identifier) is a string that follows the syntax

<scheme name> : <hierarchical part> [<query>] [# <fragment>]

Examples:

- URLs
  http://elvis.com/biography.html#Birth
- File identifiers
  file:///c:/users/elvis/tripToMoon.txt
- FTP
  ftp://elvis@nsa.gov
- Mail To
  mailto:him@elvis.com?subject=Where%20%are%20you

All URLs are URIs, but not all URIs are URLs ("dereferenceable")

Each KB & each entity has a URI

Each KB on the Semantic Web has a URI:

ElviPedia: http://elvis-alive.org/
ElviPedia': http://elvipedia.com/
ElvisKB: http://elvis.org/kb/
YAGO: http://yago-knowledge.org/

Each of them forms a namespace.

Each entity in a KB has a qualified name, which is also a URI:

URI of ElviPedia:
http://elvis.org/kb/
Name in that namespace:
Elvis
Qualified name of Elvis in ElviPedia:
http://elvis.org/kb/Elvis
(again a URI)

Each KB & each entity has a URI

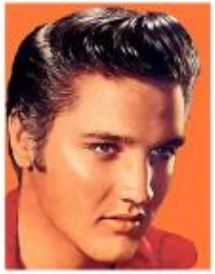
http://elvipedia.com/

http://elvipedia.com/Elvis

http://yago-knowledge.org/

http://yago-knowledge.org/Elvis

http://elvis-alive.org/



http://elvis-alive.org/Elvis

http://elvis.org/kb/

http://elvis.org/kb/Elvis

Namespaces

http://elvis.is/king/of/sing

World-wide unique mapping to domain owner

in the responsibility of the domain owner

=> There should be no overlap

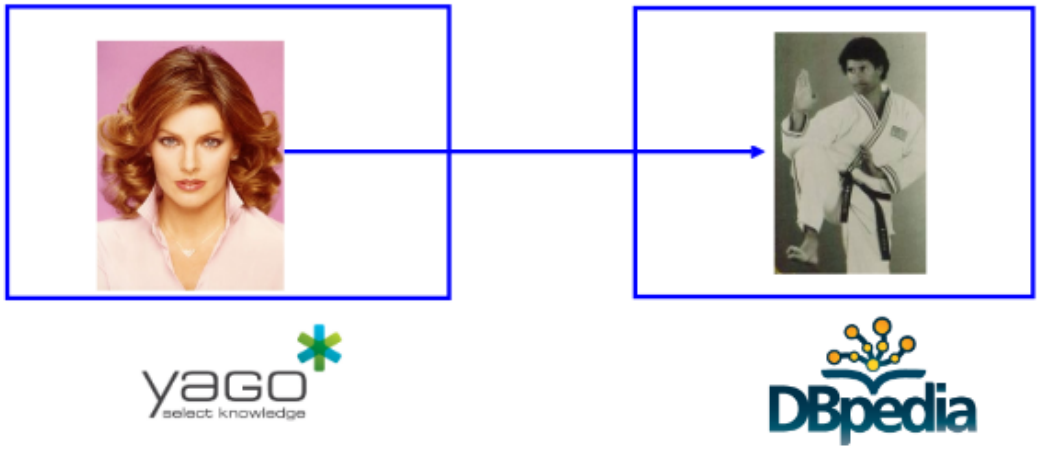
- a company can create URIs to identify its products
- an organization can assign sub-domains and each sub-domain can define URIs
- individual people can create URIs from their homepage
- people can create URIs from any URL for which they have exclusive rights to create URIs

Cross-referencing

A KB can make statements about entities defined in other KBs.

@prefix y: <http://yago-knowledge.org/>
@prefix d: <http://dbpedia.org/>

y:Priscilla y:loves d:Mike Stone .



30

Standard vocabulary

A KB can define vocabulary that is used by other KBs.

yago
select knowledge

y:Singer
- subclasses
- superclasses
- label
- ...

AlizéeKB

y:Singer

type

RDF and RDFS vocabularies

RDF is also a vocabulary (=KB) that defines basic notions of KB representation.

@prefix rdf: <http://www.w3.org/... '
rdf:type, rdf:Property, rdf:Statement .

We can use notions from this KB:

rdf:type → y:Singer

RDFS is a vocabulary (=KB) that defines basic notions for class representation.

@prefix rdfs: <http://www.w3.org/.../rdfs/>
rdfs:label, rdfs:subClassOf,
rdfs:domain, rdfs:range,
rdfs:Class, rdfs:Resource ← "entity"

y:Singer —rdfs:subClassOf→ y:Person

Sharing vocabularies

Shared vocabularies mean

- shared work in defining entities
- inter-operability of KBs

Some shared vocabularies have become standards on the Semantic Web.
They have a standard namespace prefix.

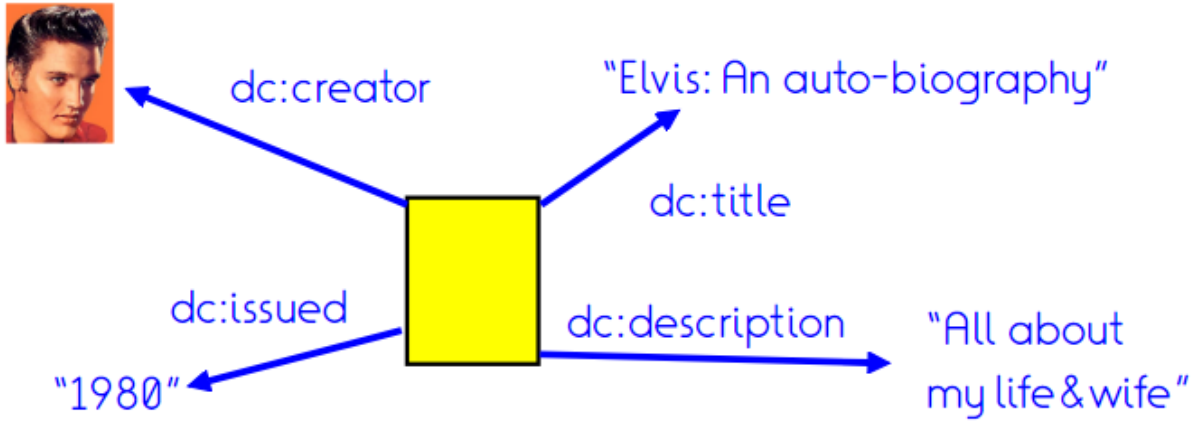
More vocabularies

- Dublin Core (for describing documents)
  http://purl.org/dc/elements/1.1/
- Schema.org (for Web content)
  http://schema.org
- Creative Commons (types of licences)
  http://creativecommons.org/ns#
- Facebook Open Graph (for Web content)
  http://ogp.me/
- FOAF (Friend of a Friend; for contact information)
  http://xmlns.com/foaf/spec/

Dublin Core

Dublin Core is a vocabulary (=KB) of terms (=entities) for describing documents.

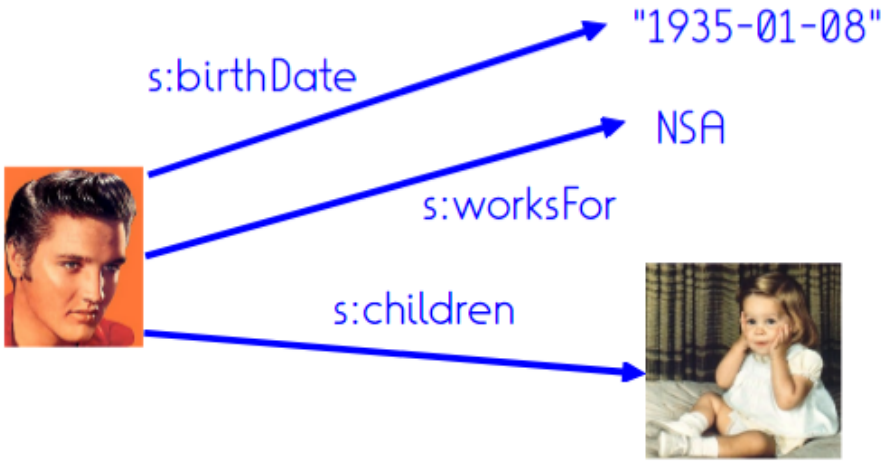
dc:creator, dc:title, dc:format, dc:MediaType, dc:language...



Schema.org

Schema.org is a KB by Google, Yahoo & Microsoft for describing Web content.

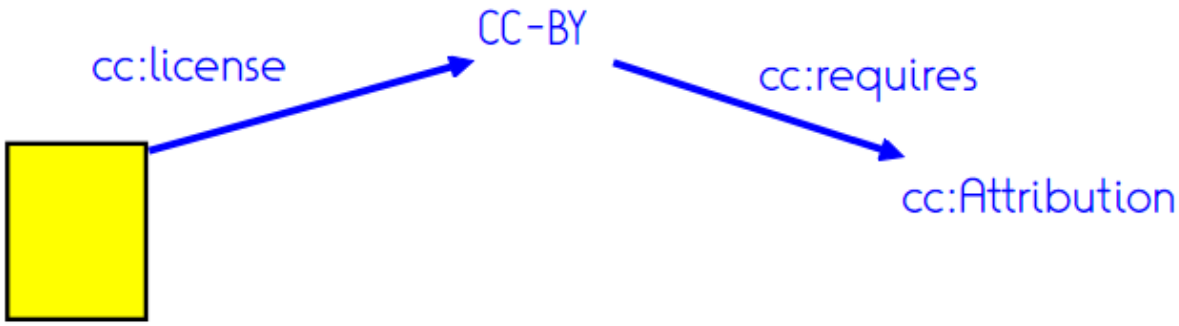
s:Person, s:Movie, s:address,
s:follows, s:worksFor, ...



Creative Commons

Creative Commons provides their vocabulary in RDF.

cc:license, cc:attributionName,
cc:permits, cc:Reproduction, ...



Def: Dereferenceable / Cool URI

A dereferenceable URI (also: Cool URI) is a URI that returns an RDF snippet if accessed on the Internet by an RDF client.

http://elvispedia.org/Elvis

```
@prefix e: <http://elvispedia.org/>
e:Elvis e:sings e:aSong .
e:Elvis e:born e:Tupelo .
...
```

Try, e.g., wget http://dbpedia.org/resource/Elvis_Presley -O elvis.rdf - -header="Accept: application/rdf+xml"

https://www.wikidata.org/wiki/Special:EntityData/Q565400.rdf

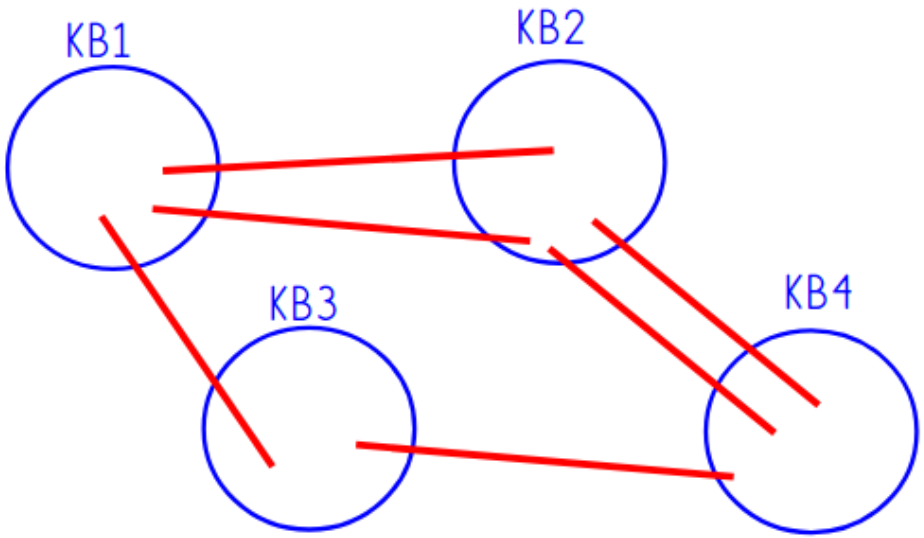
Cool URIs can be traversed

```
@prefix e: <http://elvispedia.org/>
@prefix d: <http://dbpedia.org/>
e:Priscilla e:loves d:MikeStone
...
```

http://dbpedia.org/MikeStone

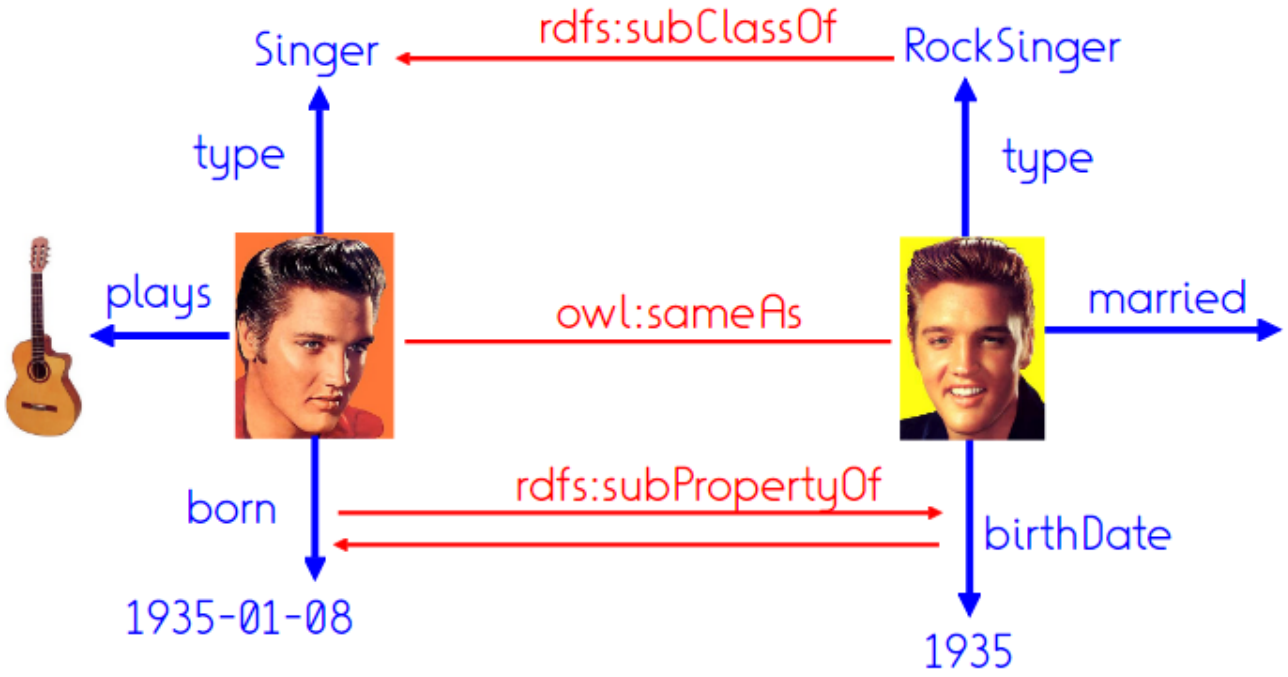
```
@prefix d: <http://dbpedia.org/>
@prefix rdf: <http://w3c.org/.../rdf>
d:MikeStone rdf:type d:KarateClown
d:MikeStone d:livesIn d:LosAngeles
...
```

Cool URIs can be traversed



The standard vocabularies (RDF, RDFS, schema.org, Creative Commons, etc.) all provide dereferenceable URIs, as do many KBs.

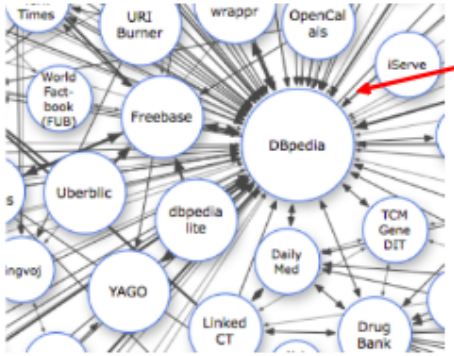
Interlinking on the Semantic Web



OWL and RDF are standard vocabularies for the linking.

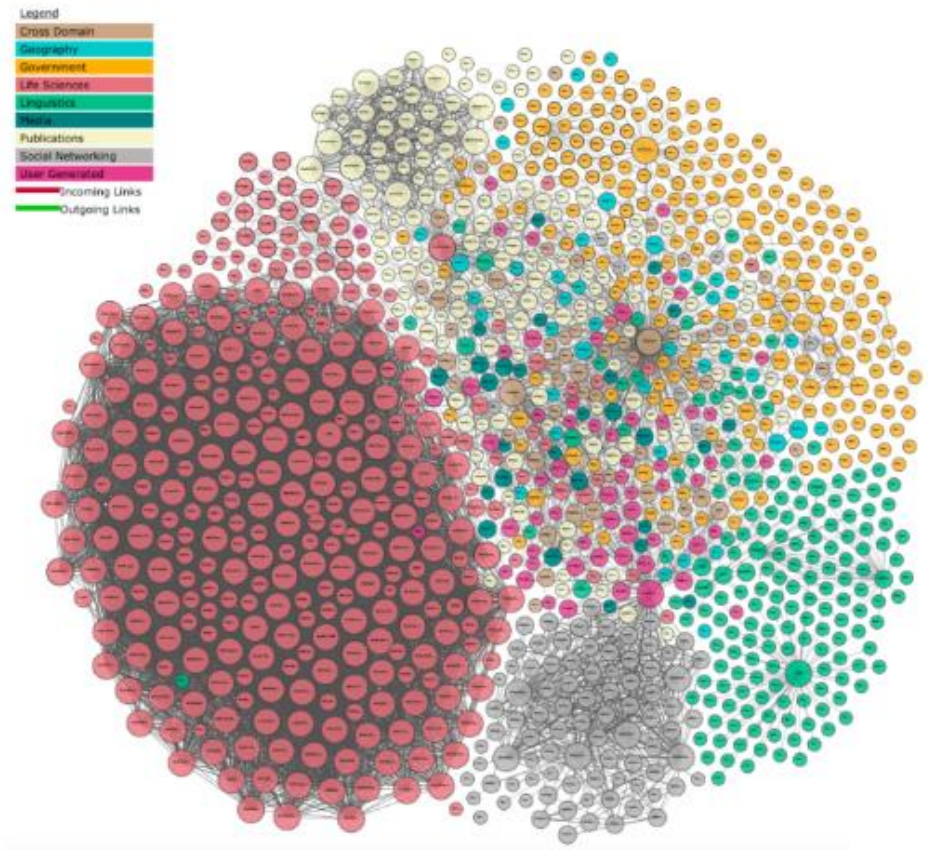
Def: Linked Open Data Project

The goal of W3C's Linked Open Data Project is to publish and link open KBs. The project links equivalent entities and equivalent relations across different KBs.



This arrow means: equivalent entities between iServe and DBpedia have been linked.

The Linked Open Data Project



As of 2017: 10,000 KBs

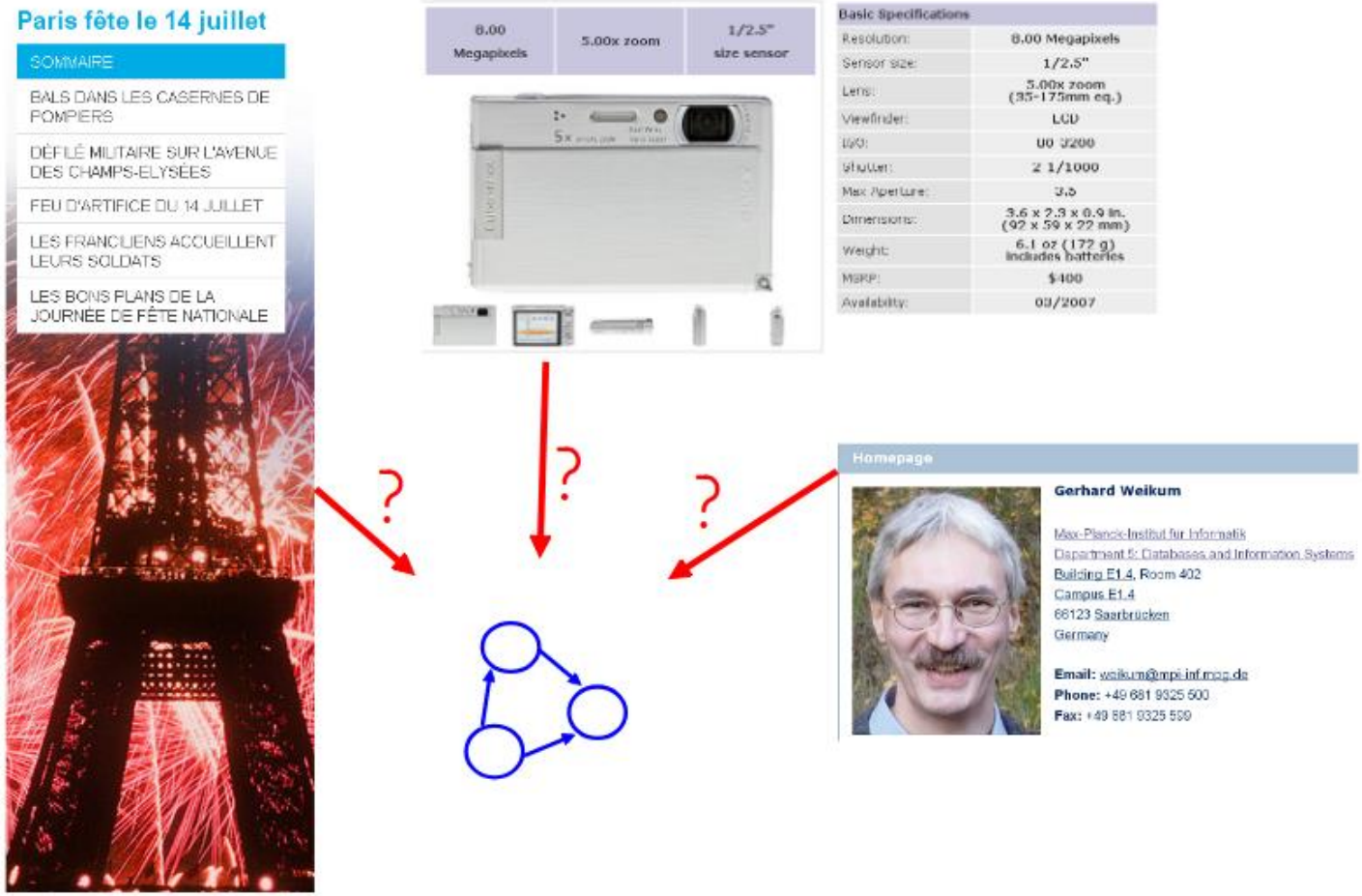
The Linked Open Data Project

Existing KBs include

- US census data
- BBC music database
- Gene ontologies
- DBpedia general knowledge, + YAGO, + Cyc etc.
- UK government data
- geographical data in abundance
- national library catalogs (USA, Germany etc.)
- publications (DBLP)
- commercial products
- all Pokemons

...and many more

How do we get HTML pages to RDF?



Defining a fact with an entity object

A tag with "property" and "resource" defines a fact between subject and URI.

```
<div vocab="http://schema.org/"
 resource="http://martin.org/me" typeOf="Person">
 <span property="name">Martin Th</span><br>
 <span property="homeLocation" resource=
 "http://yago.org/Memphis">Memphis</span>
</div>
```

<http://martin.org/me> <http://schema.org/homeLocation>
<http://yago.org/Memphis> .

RDFa Example

Contact

Fabian M. Suchanek
Département INFRES (Office C201-6),
Télécom ParisTech
46 rue Barrault
75013 Paris
France

W3C Semantic Web

RDFa Validator

https://suchanek.name/index.html#contact

```
@prefix ns1: <http://schema.org/> .
@prefix ns2: <http://www.w3.org/ns/rdfa#> .
@prefix ns3: <http://ogp.me/ns/article#> .
@prefix og: <http://ogp.me/ns#> .

<http://suchanek.name/fabian> a ns1:Person;
  og:description "full professor";
  og:image <https://suchanek.name/about/fabian.jpg>;
  og:title "Fabian M. Suchanek";
  ns1:address [ a ns1:PostalAddress;
      ns1:addressCountry <http://yago-knowledge.org/resc
      ns1:addressLocality "Paris";
      ns1:postalCode "75013";
      ns1:streetAddress "46 rue Barrault" ];
  ns1:image <https://suchanek.name/about/fabian.jpg>;
  ns1:jobTitle "full professor";
  ns1:name "Fabian M. Suchanek";
  ns1:url <https://suchanek.name>;
  ns1:worksFor <http://www.enst.fr> .
```

Summary: RDFa embeds into HTML

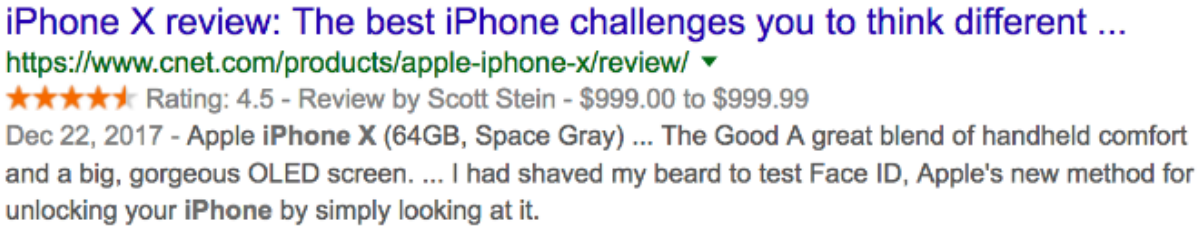
Advantages:

- Grass root appeal
  (everybody can start annotating pages)
- No data duplication
  (all data in one file)
- Publisher independence
  (everybody can use his own attributes)

Standards that are similar to RDFa are

- Microformats
- Microdata
- JSON-LD

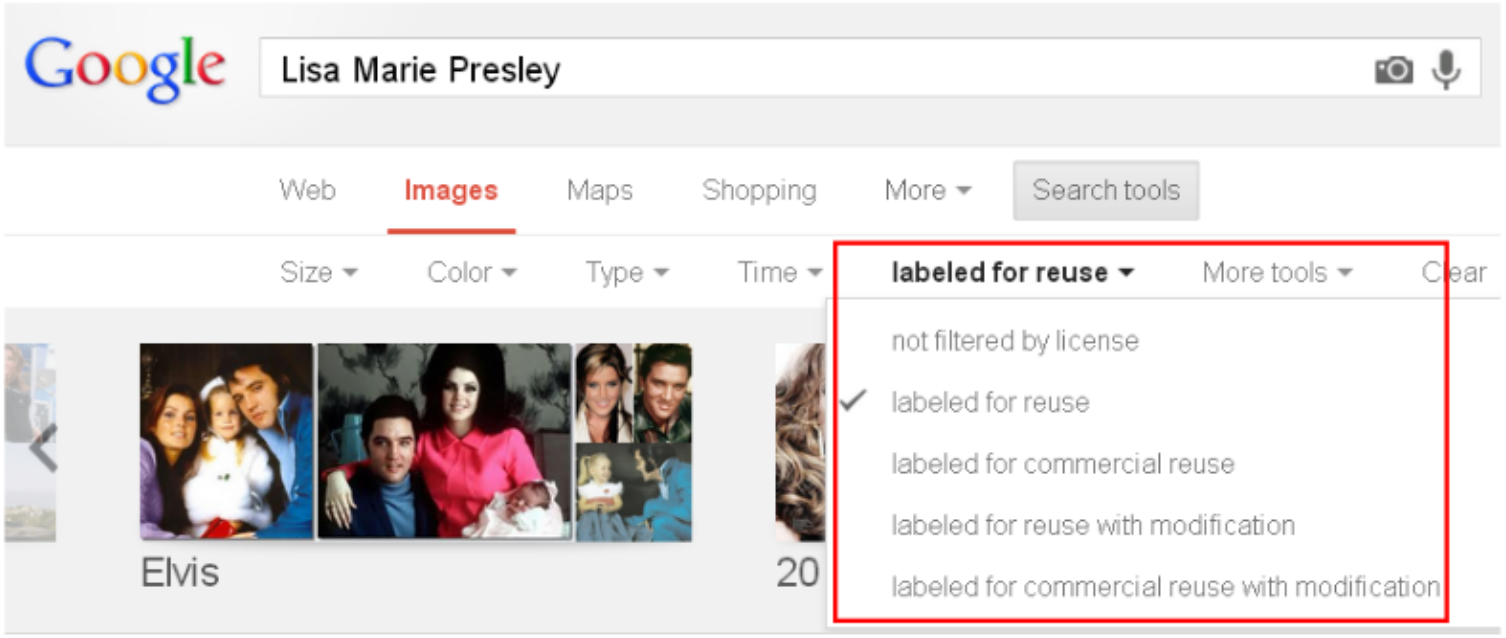
Search engines scrape RDFa&JSON-LD



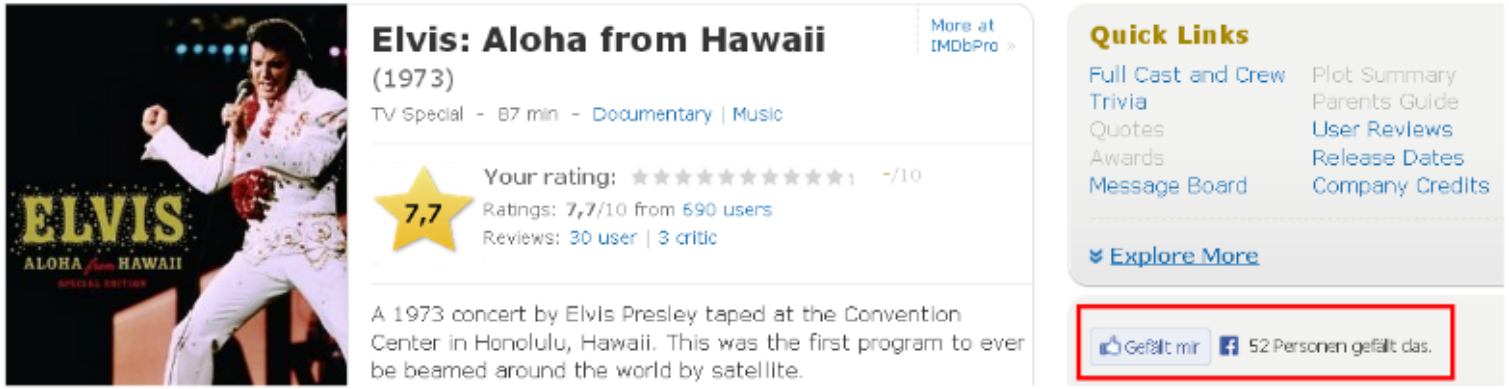
JSON-LD embedded in Web page:

```
<script type="application/ld+json">
  {
    "@context":   "http://schema.org",
    "@type":     "Product",
    "name":      "Apple iPhone X",
    "description": "iPhone X is an overdue and winning evolution of the iPh
    "image":     "https://cnet1.cbsistatic.com/img/ZQICw4aW2fNpbmN34
    "brand": {
            "@type": "Thing",
```

Search engines read licenses



Facebook Like Button uses RDFa



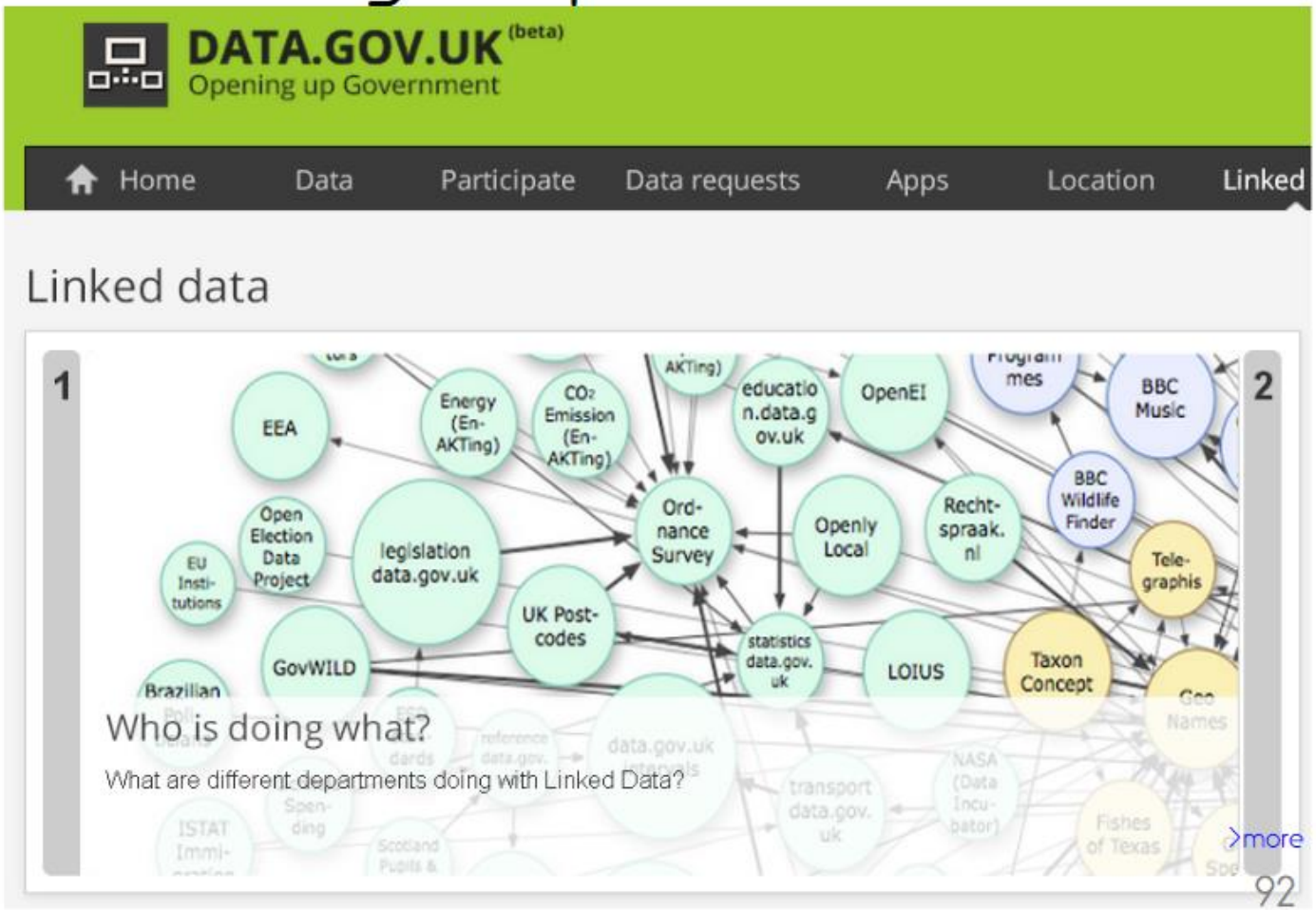
@prefix og: <http://ogp.me/ns#> .

<http://www.imdb.com/title/tt0167923/?ref=fnaltt2> og:description
"A 1973 concert by Elvis Presley taped in Honolulu, Hawaii";
og:sitename "IMDb";
og:title "Elvis: Aloha from Hawaii (1973)";
og:type "video.tv-show";

Facebook public pages have JSON-LD


```
<script type="application/ld+json">
{"@context":"http://schema.org",
 "@type":"Organization",
  "name":"ELVIS PRESLEY", ...
```

UK and US govts publish RDF



References

- Selected references

F. Suchanek, G. Kasneci, G. Weikum:
"Yago: a core of semantic knowledge", WWW 2007

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak:
"Dbpedia: A nucleus for a web of open data", ISWC 2007

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., Tom M. Mitchell:
"Toward an Architecture for Never-Ending Language Learning" (NELL), AAAI 2010

R. Navigli, S. Ponzetto:
"BabelNet: The automatic construction [of a] multilingual semantic network", Journal of AI 2012

D. Vrandecic, M. Krötzsch:
"Wikidata: a free collaborative knowledgebase", Comm. of ACM 2014

- Further reading
  - qa.mpi-inf.mpg.de

- Slides
  - Adapted from Fabian Suchanek and Rishiraj Saha Roy

~~Assignment 9~~

- No assignment ☺
- Tutorial today: Exam questions

Take home

- IE important tool for building structured knowledge
- Wikipedia popular resource
- Free text extraction harder but possible
- KBs in widespread use in tech companies
  - Actual methods guarded secrets
  - Source of data not always known
- Signature application: Question answering
  - Challenge: From unstructured user question to structured KB query
- Semantic web: Vision of interlinked and machine-readable internet
  - Schema reuse essential for (simple) machine-readability