Automated knowledge base construction
1. Introduction

Simon Razniewski

Summer term 2022

Outline

1. **Introducing each other**
2. Course organization
3. What, Why, How
4. Lab 1

Simon Razniewski

- Senior Researcher at MPII, Department 5
  - Heading “Knowledge Base Construction and Quality” area

- Background
  - Assistant professor FU Bozen-Bolzano, Italy, 2014-2017
  - PhD FU Bozen-Bolzano, 2014
  - Diplom at TU Dresden, 2010

- Research areas:
  - Logics, databases, Semantic Web
  - More recently IR, (applied) NLP, ML, …

- Research focus: Knowledge base construction and quality
  - Analyzing what knowledge bases know, and what they don’t
  - Commonsense knowledge base construction

Tutorial teachers

- Hiba Arnaout
- Shrestha Ghosh
- Sneha Singhania
- Tuan-Phong Nguyen

- Doctoral researchers at D5, MPII
- Knowledge base construction, question answering, knowledge coverage, commonsense knowledge, …

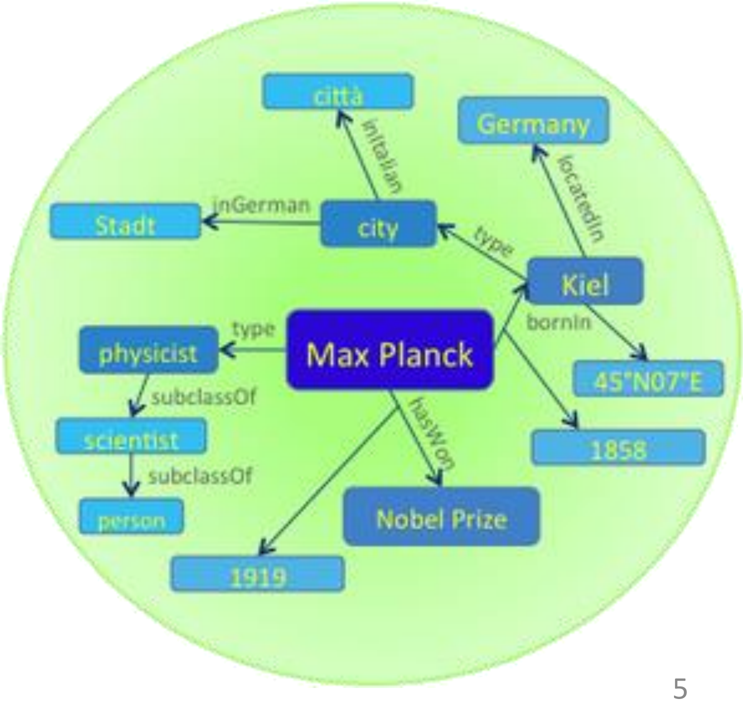
Department 5

- Department 5: Database and information systems

- Knowledge discovery: *extracting*, *organizing*, *searching*, *exploring* and *ranking* facts from *structured*, *semi-structured*, *textual* and *multimodal* information sources

- YAGO select knowledge Knowledge Base
  - Earliest prominent machine-generated knowledge base (2007)
  - Contains more than 10 million entities and more than 120 million facts

- Gerhard Weikum 259th most cited computer scientist worldwide



And you?

- Course of study
- Preknowledge
- …
- Comments?

- https://tinyurl.com/4xpk8enh

Outline

1. Introducing each other
2. **Course organization**
3. What, Why, How
4. Lab 1

Learning outcomes

- Knowledge
  - What AKBC is about (“What”)
  - What AKBC is good for (“Why”)
  - What main tasks and challenges in AKBC are
  - What common approaches to problems in AKBC are (“How”)

- Skills
  - Analyze potentials and limitations of AKBC approaches
  - Learn to choose right source and method for right task
  - Implement simple solutions for main problems in AKBC
    - Scraping, typing, linking, …

- Abilities
  - Build your own AKBC pipeline for a problem

→ Very practical focus!

Prerequisites

- Basic concepts of ML
  - We won’t go deep
- Python programming
  - Essential
  - Still time to learn

- Helpful but not required
  - Basic notions of information retrieval (IRDM?)
  - Computational linguistics (SNLP?)

Formal organization

- Credit points: 6, hours: 180 (!)
- Registration
  - Subscribe to the mailing list https://groups.google.com/g/akbc2022/
  - Register in HISPOS until 4.7. for the exam
- When?
  - Lecture: Wednesday 12:15-13:45
  - Lab: Wednesday 16:15-17:45
- How to pass this course?
  - 8 small practical assignments
    - Pass/fail
    - To be admitted to exam, pass at least 6
  - Oral exam

Assignments

- Published on lecture day (Wednesday)
- Due Monday 23:59 the week after

- Labs are there to start solving the assignments

- Discussing assignments together is allowed, but **each student must write their own solution**
  - No sharing of code!
  - Plagiarism = course failed for both
  - Avoid triangular plagiarism = cite sources
    - *"Approach for NER adapted from stackoverflow.com/how-to-…"*

- Libraries that solve core tasks not allowed
  - In doubt ask..

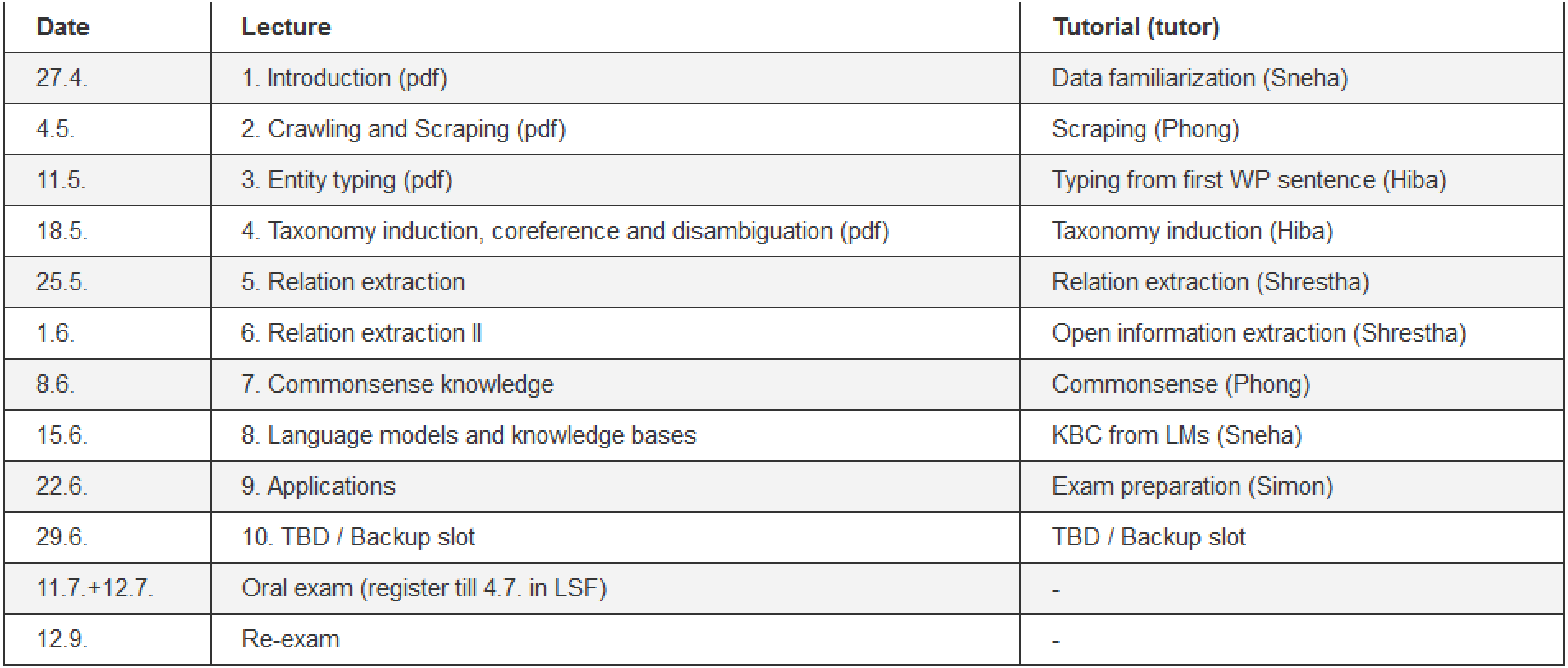
- Weekly assignments are evil!?
  - Established psychological "trick" to help you learn and pass!

Assignment content

- Coding
- 4 assignments are in competition format
  - Crisp input/output problem specification
    - “From the first sentence of Wikipedia, extract the type of an entity”
  - Labelled training/test data set
  - Unseen (hidden) evaluation dataset
    - To avoid overfitting

  → Ranked list by a standard metric, e.g., precision or F1-score
    - But pass/fail does not depend on rank

Schedule



Date	Lecture	Tutorial (tutor)
27.4.	1. Introduction (pdf)	Data familiarization (Sneha)
4.5.	2. Crawling and Scraping (pdf)	Scraping (Phong)
11.5.	3. Entity typing (pdf)	Typing from first WP sentence (Hiba)
18.5.	4. Taxonomy induction, coreference and disambiguation (pdf)	Taxonomy induction (Hiba)
25.5.	5. Relation extraction	Relation extraction (Shrestha)
1.6.	6. Relation extraction II	Open information extraction (Shrestha)
8.6.	7. Commonsense knowledge	Commonsense (Phong)
15.6.	8. Language models and knowledge bases	KBC from LMs (Sneha)
22.6.	9. Applications	Exam preparation (Simon)
29.6.	10. TBD / Backup slot	TBD / Backup slot
11.7.+12.7.	Oral exam (register till 4.7. in LSF)	-
12.9.	Re-exam	-

Outline

1. Introducing each other
2. Course organization
3. **What, Why, How**
4. Lab 1

3. Introduction to AKBC

I. **Motivation**

II. Terminology

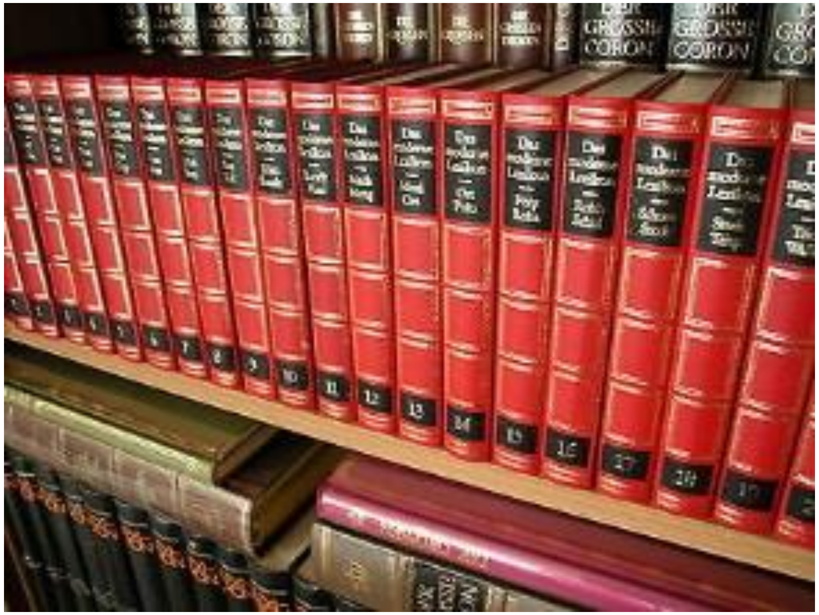
III. Topics

IV. Construction techniques

V. Applications

VI. Past, present and future

I. Motivation



- https://en.wikipedia.org/wiki/Max_Planck_Institute_for_Informatics

WIKIDATA

- https://www.wikidata.org/wiki/Q565400

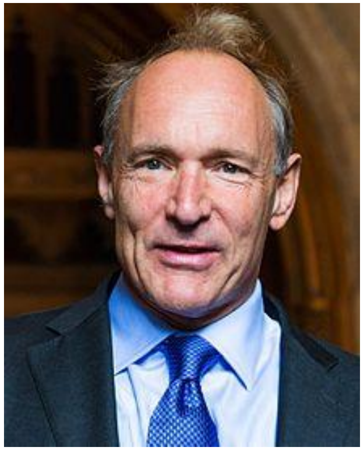
What for?

- One central hub for interlanguage interlinking of 100+ Wikipedia editions
- Your AI chatbot wants to know where MPII, MIT and KAIST are located? → structured query
- A library wants to distinguish which of the 100+ literary John Smiths wrote *“A description of New England”*? → Wikidata ID

Samples of advanced queries

- Who discovered the most planets: http://tinyurl.com/y7rldyqc
- Distribution of places ending with “-weiler” in Germany: https://w.wiki/67o
- Living relatives of Louis XIV of France: https://w.wiki/549E

The Semantic Web



- Term coined by Tim Berners-Lee for a machine-readable Web

- Web content originally from humans for humans

→ Make machines read human language, or make humans write machine-readable structured data?

3. Introduction to AKBC

I. Motivation

II. Terminology

III. Topics

IV. Construction techniques

V. Applications

VI. Past, present and future

Facts (triples) and their constituents

- **Entities**: Objects about which statements can be made
  Paris; Trump; Irony
- **Property**/predicate/relation/attribute: What can be said
  locatedIn(entity, location), worksAt(person, organization), antonymOf(term, term)
- **Fact**/statement/claim/triple: Core building block of KBs
  <Paris, locatedIn, France>

→General form:

<subject, predicate, object>

<s, p, o>

Subjects and objects

- Machine-generated identifiers
  - Wikidata: *Q4262, Q67245*
- Canonical name strings
  - DBpedia, YAGO: *“John_Smith_(politician)”*
- Internationalized resource identifier (IRI)
  - Semantic web: *http://dbpedia.org/resource/Max_Planck*
- General phrases
  - TupleKB: *<industry, grow over, past few decade>*
- Literals: Attribute values that are no entities
  - *www.mpi-inf.mpg.de*
  - Often with units: *1.63m; 54.85° N*

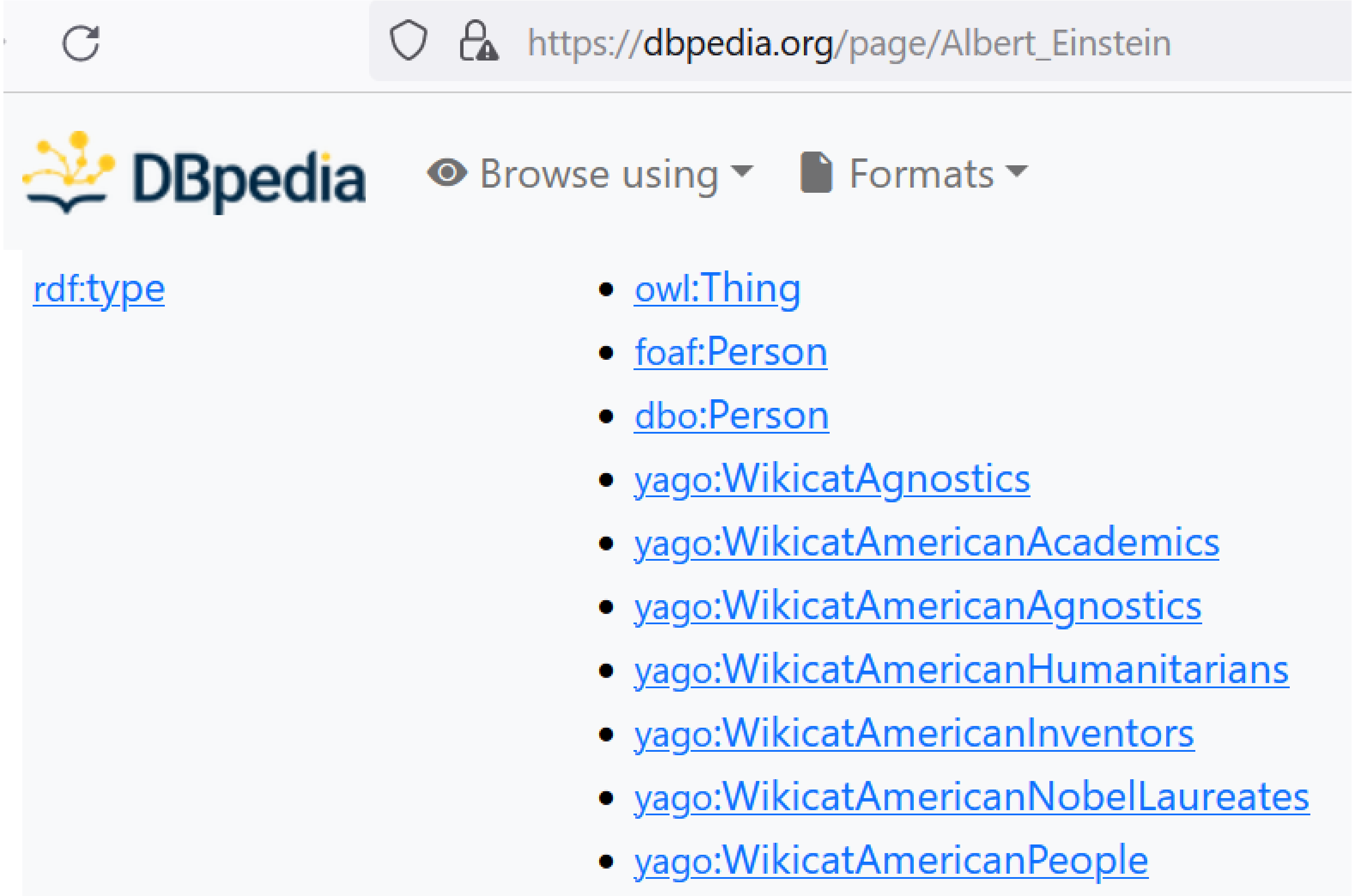
- Same for predicates, sometimes canonicalized, sometimes just text

Classes and class hierarchies

- **Classes**/types: Allow to group similar entities
  Presidents, nouns, Greek gods

- **Type/property hierarchy**: Tree-like hierarchy among types/properties (cf. inheritance in object-oriented programming)
  <Town, subclassOf, Administrative_unit>

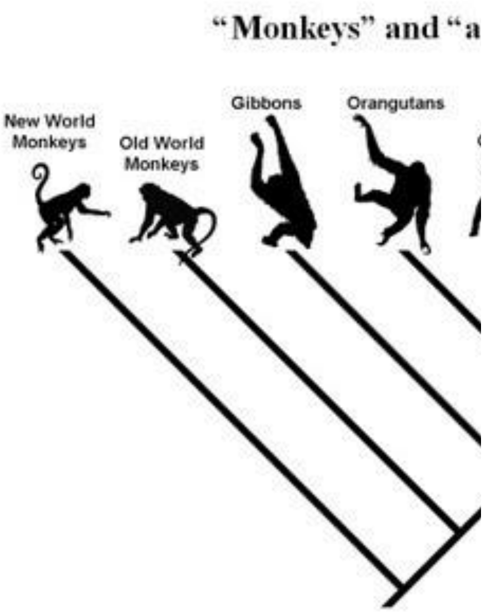
Classes

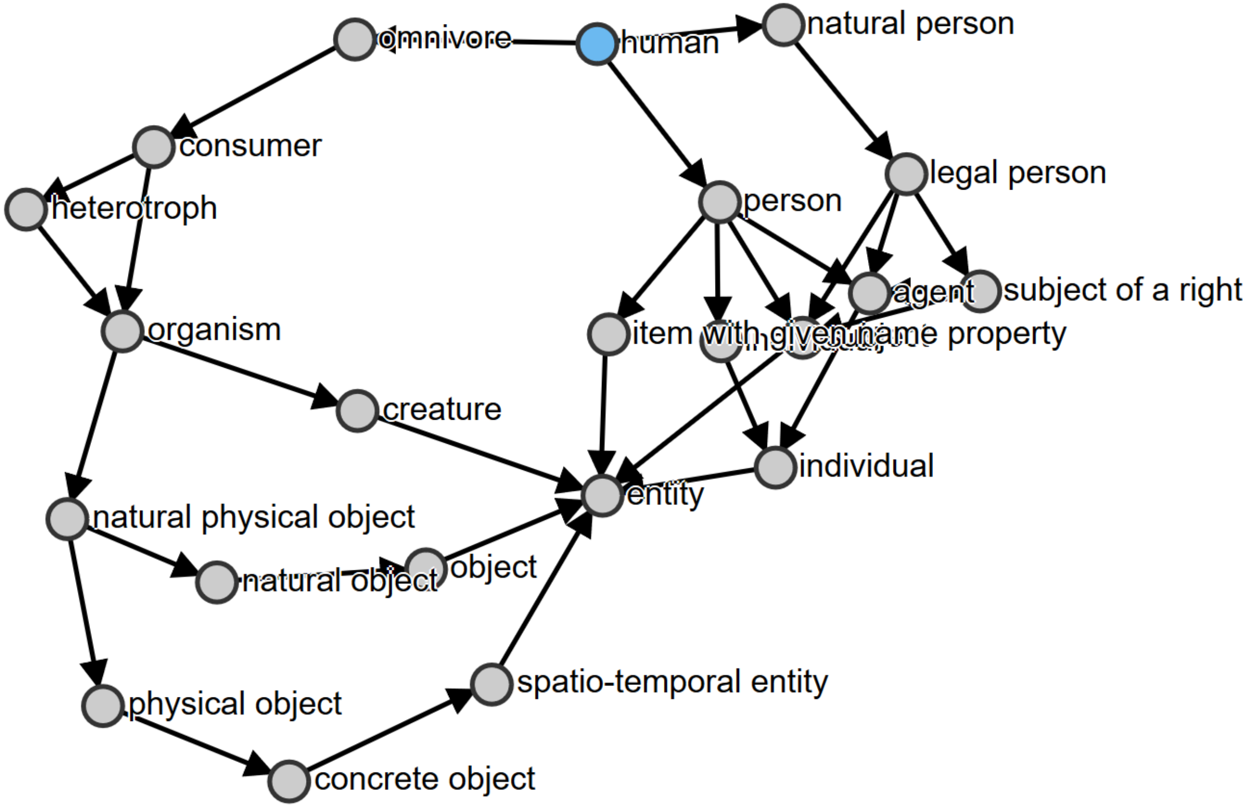


rdf:type

- owl:Thing
- foaf:Person
- dbo:Person
- yago:WikicatAgnostics
- yago:WikicatAmericanAcademics
- yago:WikicatAmericanAgnostics
- yago:WikicatAmericanHumanitarians
- yago:WikicatAmericanInventors
- yago:WikicatAmericanNobelLaureates
- yago:WikicatAmericanPeople

Taxonomies





https://angryloki.github.io/wikidata-graph-builder/?property=P279&item=Q5

Knowledge base: Definition

A knowledge base (KB) is a collection of structured data about entities and relations with the following characteristics:

- Content: The KB contains entities and their semantic types for a given domain of interest. Additionally, attributes of entities (including numeric and string literals) and relationships between entities are captured.
- Schema and Scale: Unlike a conventional database, there is often no pre-determined relational schema where all knowledge has to fit into a static set of relations. If fixed, longitudinal evolution must allow ad-hoc additions where the set of types and relations may grow to ten or hundred thousands.
- Open Coverage: New entities and facts emerge and get covered in new web sources at high rate. Therefore, we have to view KB construction and maintenance as a "never-ending" task, following an open world assumption and acknowledging the high pace of real-world changes.

[Weikum et al., FnT 2021]

3. Introduction to AKBC

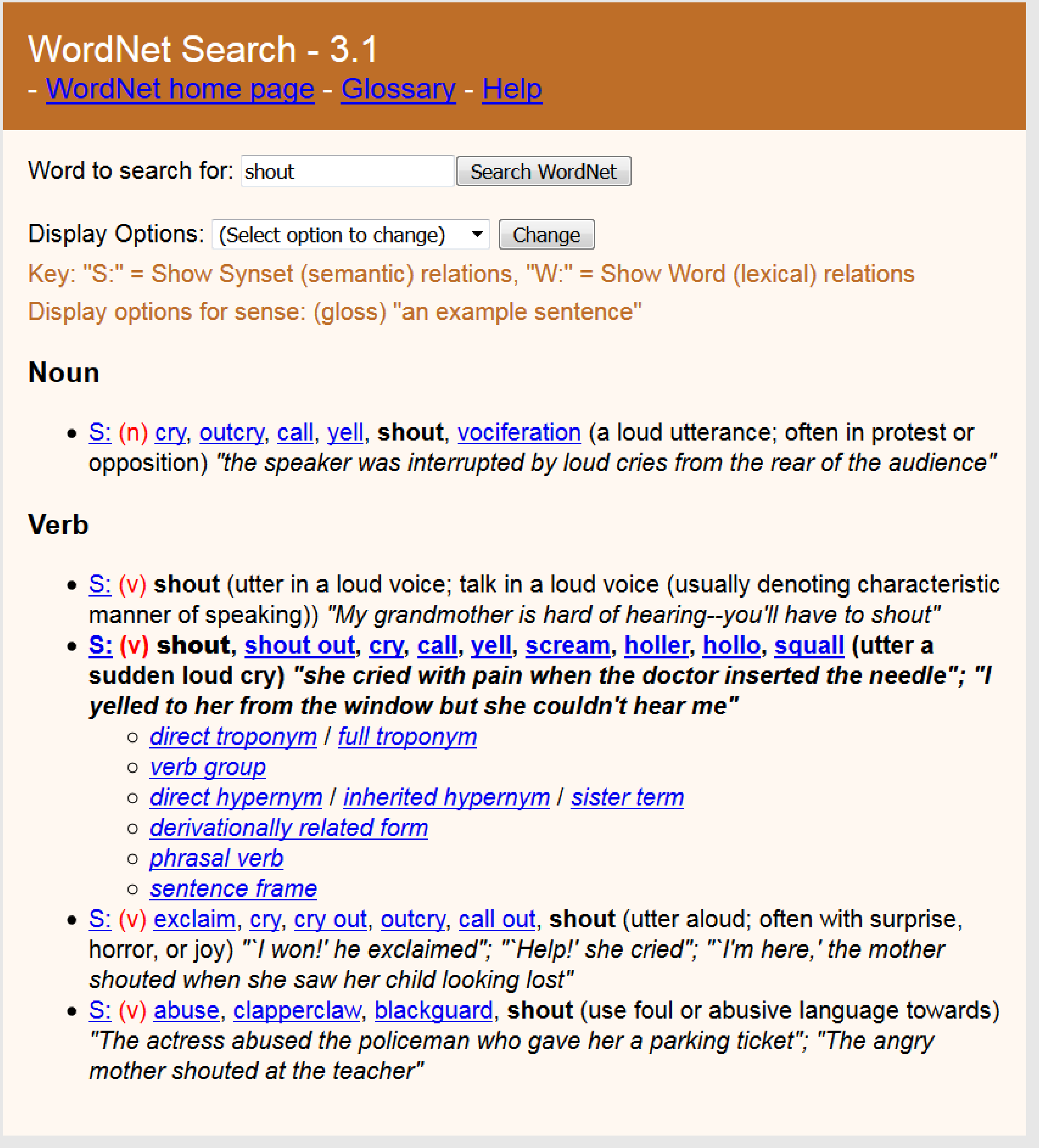
I. Motivation
II. Terminology
III. Topics
IV. Construction techniques
V. Applications
VI. Past, present and future

Common topics of knowledge bases

- Lexical knowledge
  - *<shout, isA, verb>*
  - *<shout, subformOf, communicate>*
- Instance knowledge ("Encyclopedic KBs"):
  - *<Paris, capitalOf, France>*
  - *<MPII, foundedIn, 1988>*
  - *<Angela Merkel, major, Physics>*
- Class knowledge ("Commonsense"):
  - *<Pizza, is, tasty>*
  - *<Elephant, color, grey>*
  - *<turnOnPC, requires, power>*

Lexical KBs

- WordNet (1995)
- FrameNet (1998)
- (Wiktionary (2002))
- SenticNet (2010)
- …



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: shout

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) cry, outcry, call, yell, **shout**, vociferation (a loud utterance; often in protest or opposition) *"the speaker was interrupted by loud cries from the rear of the audience"*

Verb

- S: (v) **shout** (utter in a loud voice; talk in a loud voice (usually denoting characteristic manner of speaking)) *"My grandmother is hard of hearing--you'll have to shout"*
- **S: (v) shout, shout out, cry, call, yell, scream, holler, hollo, squall (utter a sudden loud cry) *"she cried with pain when the doctor inserted the needle"; "I yelled to her from the window but she couldn't hear me"***
  - *direct troponym* / *full troponym*
  - *verb group*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*
  - *phrasal verb*
  - *sentence frame*
- S: (v) exclaim, cry, cry out, outcry, call out, **shout** (utter aloud; often with surprise, horror, or joy) *"`I won!' he exclaimed"; "`Help!' she cried"; "`I'm here,' the mother shouted when she saw her child looking lost"*
- S: (v) abuse, clapperclaw, blackguard, **shout** (use foul or abusive language towards) *"The actress abused the policeman who gave her a parking ticket"; "The angry mother shouted at the teacher"*

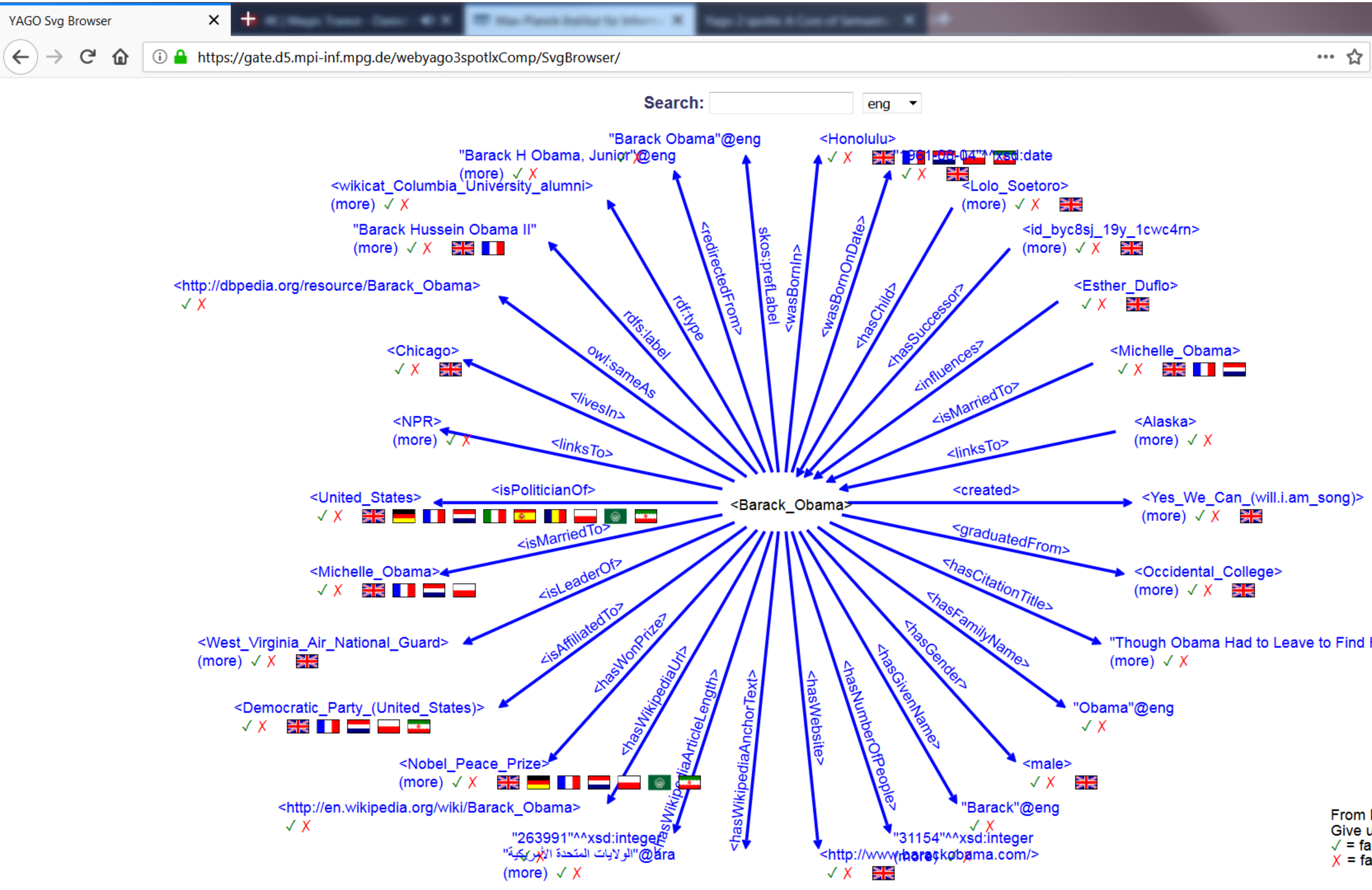
FrameNet

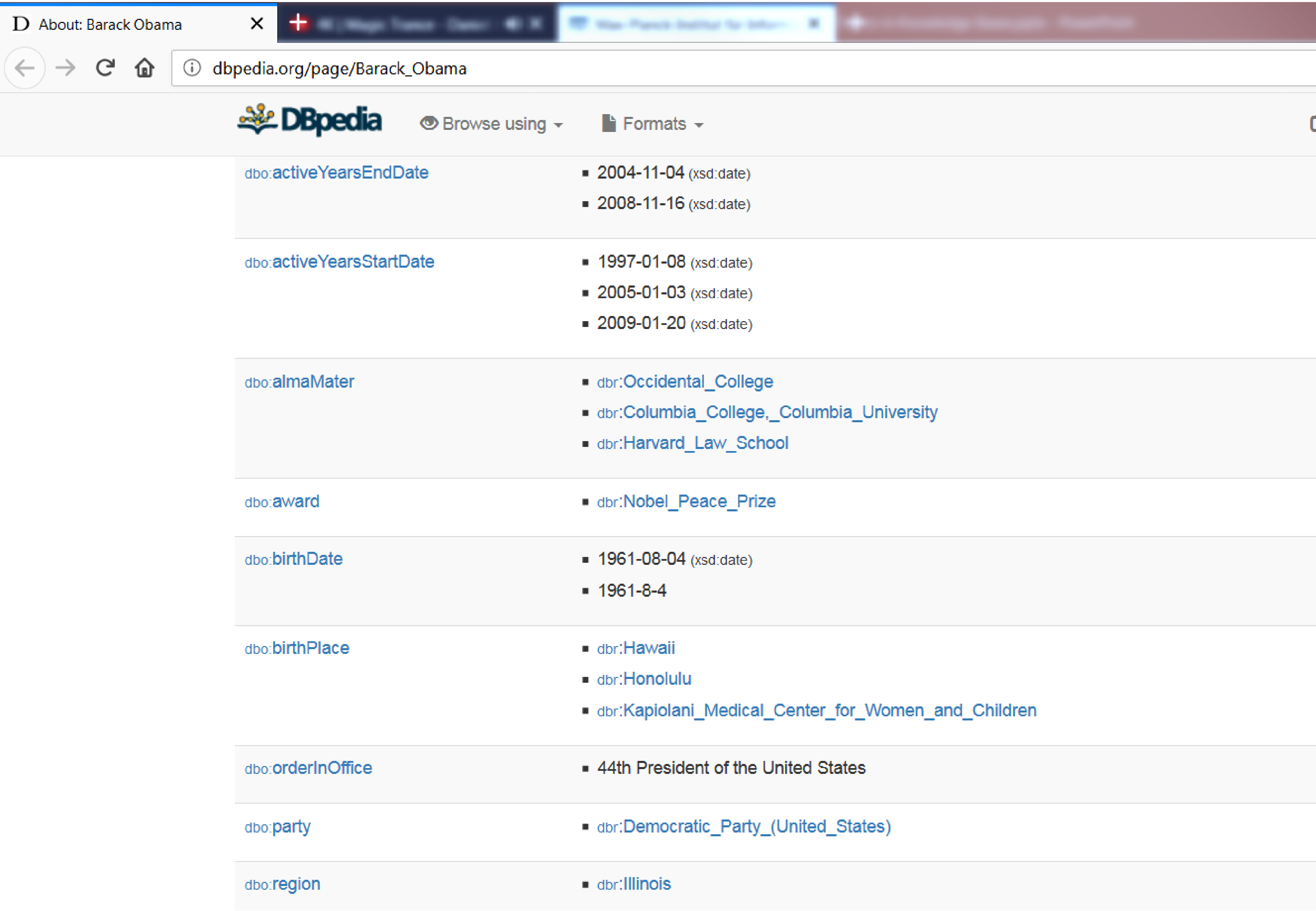
- Example Frame – “Revenge”: Because of some **injury** to something-or-someone important to an **avenger** (maybe himself), the **avenger** inflicts a **punishment** on the **offender**. The **offender** is the person responsible for the **injury**.
- Frame elements:
  - **avenger, offender, injury, injured_party, punishment.**
- Invoking terms:
  - Nouns: *revenge, vengeance, reprisal, retaliation*
  - Verbs: *avenge, revenge, retaliate (against), get back (at), get even (with), pay back*
  - Adjectives: *vengeful, vindictive*

Encyclopedic KBs (“Instance-oriented KBs”)

- Cyc (1984)
- YAGO (2007)*
- DBpedia (2007)
- Wikidata (2012)

** developed at MPII*


YAGO Svg Browser
https://gate.d5.mpi-inf.mpg.de/webyago3spotlxComp/SvgBrowser/
Search:
eng
<Barack_Obama>
"Barack Obama"@eng
"Barack H Obama, Junior"@eng
(more) ✓ ✗
<Honolulu>
✓ ✗
"1961-08-04"^^xsd:date
✓ ✗
<wikicat_Columbia_University_alumni>
(more) ✓ ✗
<Lolo_Soetoro>
(more) ✓ ✗
"Barack Hussein Obama II"
(more) ✓ ✗
<id_byc8sj_19y_1cwc4rn>
(more) ✓ ✗
<http://dbpedia.org/resource/Barack_Obama>
✓ ✗
<Esther_Duflo>
✓ ✗
<Chicago>
✓ ✗
<Michelle_Obama>
✓ ✗
<NPR>
(more) ✓ ✗
<Alaska>
(more) ✓ ✗
<United_States>
✓ ✗
<Yes_We_Can_(will.i.am_song)>
(more) ✓ ✗
<Michelle_Obama>
✓ ✗
<Occidental_College>
(more) ✓ ✗
<West_Virginia_Air_National_Guard>
(more) ✓ ✗
"Though Obama Had to Leave to Find
(more) ✓ ✗
<Democratic_Party_(United_States)>
✓ ✗
"Obama"@eng
✓ ✗
<Nobel_Peace_Prize>
(more) ✓ ✗
<male>
✓ ✗
<http://en.wikipedia.org/wiki/Barack_Obama>
✓ ✗
"Barack"@eng
✓ ✗
"263991"^^xsd:integer
"الولايات المتحدة الأمريكية"@ara
(more) ✓ ✗
"31154"^^xsd:integer
<http://www.barackobama.com/>
(more) ✓ ✗
✓ ✗
<redirectedFrom>
skos:prefLabel
<wasBornIn>
<wasBornOnDate>
<hasChild>
<hasSuccessor>
<influences>
<isMarriedTo>
<linksTo>
<created>
rdf:type
rdfs:label
owl:sameAs
<livesIn>
<linksTo>
<isPoliticianOf>
<isMarriedTo>
<isLeaderOf>
<isAffiliatedTo>
<hasWonPrize>
<hasWikipediaUrl>
<hasWikipediaArticleLength>
<hasWikipediaAnchorText>
<hasWebsite>
<hasNumberOfPeople>
<hasGivenName>
<hasGender>
<hasFamilyName>
<hasCitationTitle>
<graduatedFrom>
From I
Give u
✓ = fa
✗ = fa



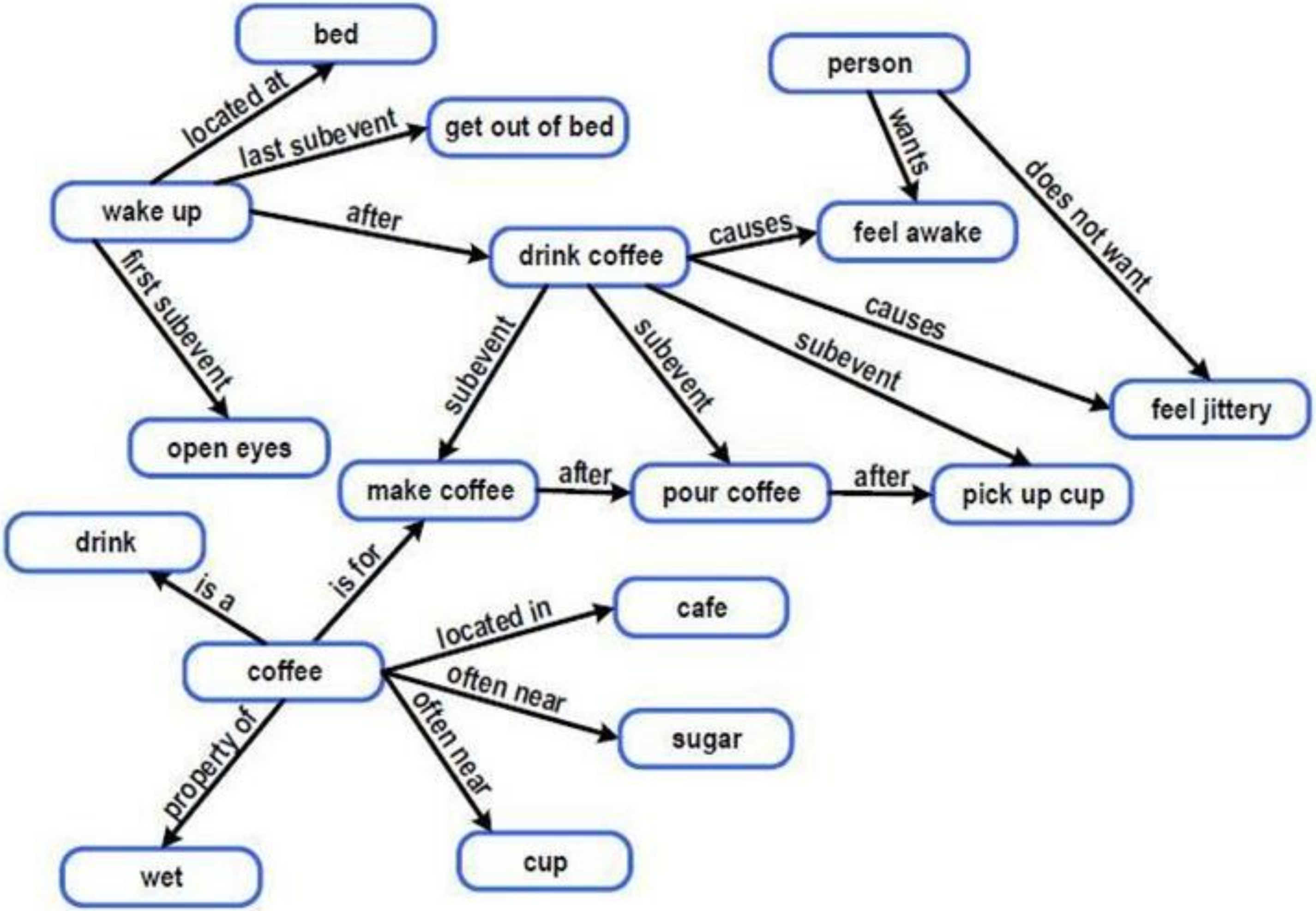
dbo:activeYearsEndDate	▪ 2004-11-04 (xsd:date) ▪ 2008-11-16 (xsd:date)
dbo:activeYearsStartDate	▪ 1997-01-08 (xsd:date) ▪ 2005-01-03 (xsd:date) ▪ 2009-01-20 (xsd:date)
dbo:almaMater	▪ dbr:Occidental_College ▪ dbr:Columbia_College,_Columbia_University ▪ dbr:Harvard_Law_School
dbo:award	▪ dbr:Nobel_Peace_Prize
dbo:birthDate	▪ 1961-08-04 (xsd:date) ▪ 1961-8-4
dbo:birthPlace	▪ dbr:Hawaii ▪ dbr:Honolulu ▪ dbr:Kapiolani_Medical_Center_for_Women_and_Children
dbo:orderInOffice	▪ 44th President of the United States
dbo:party	▪ dbr:Democratic_Party_(United_States)
dbo:region	▪ dbr:Illinois

Commonsense KBs (class-oriented)

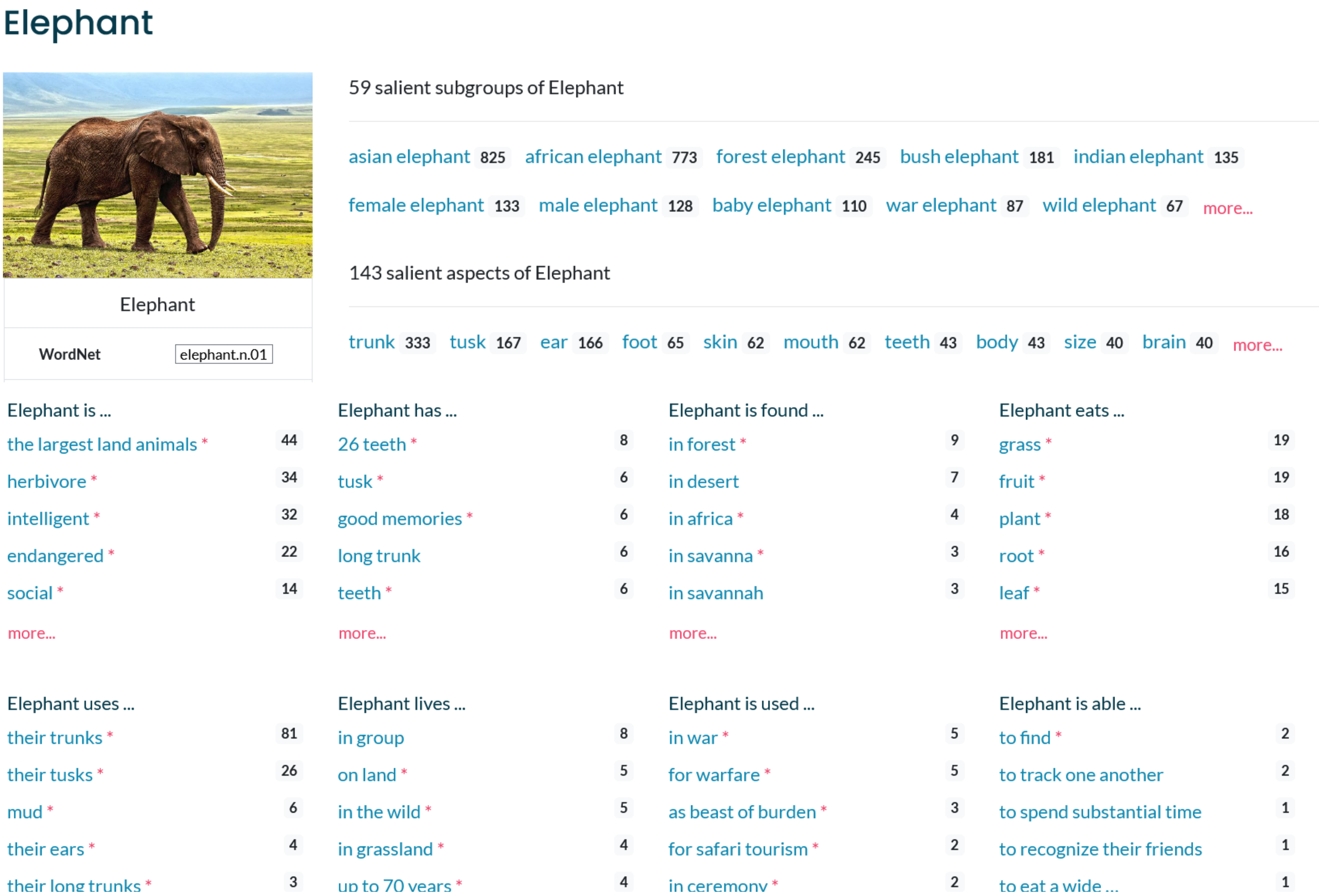
- Cyc (1984)
- ConceptNet (1999)
- TupleKB (2017)
- Quasimodo (2019)*
- Ascent (2021)*

** Developed at MPII*

ConceptNet



bed
person
located at
last subevent
get out of bed
wants
does not want
wake up
after
drink coffee
causes
feel awake
causes
first subevent
subevent
subevent
subevent
feel jittery
open eyes
make coffee
after
pour coffee
after
pick up cup
drink
is a
is for
located in
cafe
coffee
often near
often near
sugar
property of
wet
cup



Elephant

Elephant

WordNet	elephant.n.01

59 salient subgroups of Elephant

asian elephant	825
african elephant	773
forest elephant	245
bush elephant	181
indian elephant	135
female elephant	133
male elephant	128
baby elephant	110
war elephant	87
wild elephant	67

more...

143 salient aspects of Elephant

trunk	333
tusk	167
ear	166
foot	65
skin	62
mouth	62
teeth	43
body	43
size	40
brain	40

more...

Elephant is ...	
the largest land animals *	44
herbivore *	34
intelligent *	32
endangered *	22
social *	14

more...

Elephant has ...	
26 teeth *	8
tusk *	6
good memories *	6
long trunk	6
teeth *	6

more...

Elephant is found ...	
in forest *	9
in desert	7
in africa *	4
in savanna *	3
in savannah	3

more...

Elephant eats ...	
grass *	19
fruit *	19
plant *	18
root *	16
leaf *	15

more...

Elephant uses ...	
their trunks *	81
their tusks *	26
mud *	6
their ears *	4
their long trunks *	3

Elephant lives ...	
in group	8
on land *	5
in the wild *	5
in grassland *	4
up to 70 years *	4

Elephant is used ...	
in war *	5
for warfare *	5
as beast of burden *	3
for safari tourism *	2
in ceremony *	2

Elephant is able ...	
to find *	2
to track one another	2
to spend substantial time	1
to recognize their friends	1
to eat a wide ...	1

3. Introduction to AKBC

I. Motivation

II. Terminology

III. Topics

IV. Construction techniques

V. Applications

VI. Past, present and future

How to build KBs?

Possible approaches

A. Humans (CYC, ConceptNet, Wikidata)

B. Structured extraction (YAGO, DBpedia)

C. Text extraction (NELL, Textrunner)

D. Constraints and pattern mining

Our focus

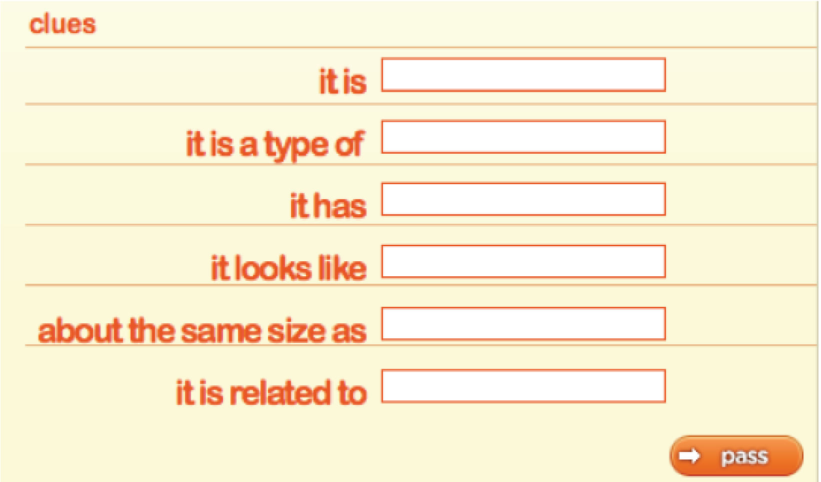
A. Humans: Experts

- Potentially best quality

- Difficult to scale
  - CYC: “In 1986, Doug Lenat estimated the effort to complete the KB to be 250,000 rules and 350 man-years of effort.”

Humans: Crowdsourcing/Gamification

- Make work fun (?)



clues

it is

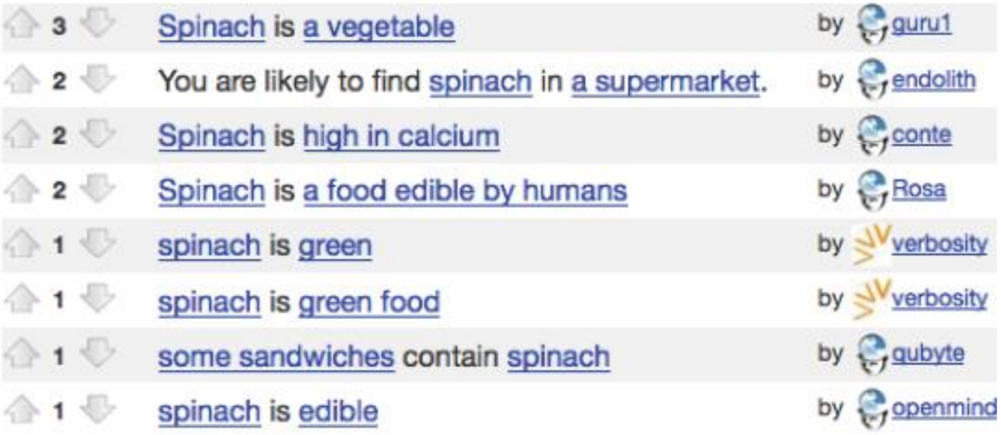
it is a type of

it has

it looks like

about the same size as

it is related to



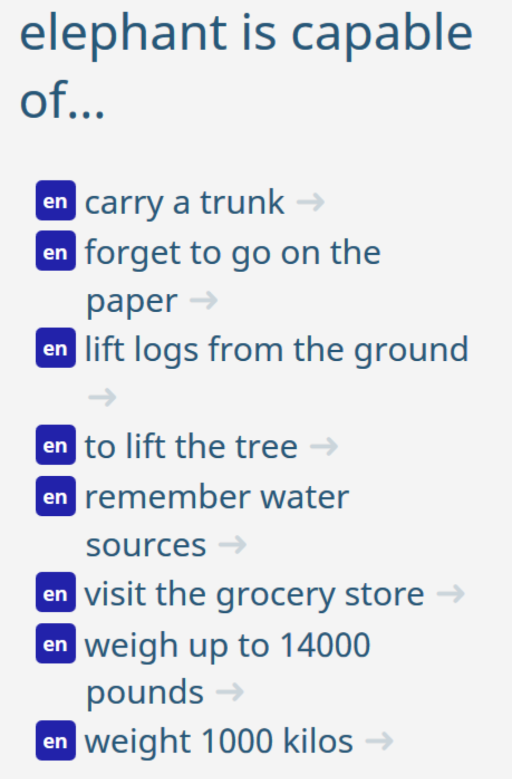
3	Spinach is a vegetable	by	guru1
2	You are likely to find spinach in a supermarket.	by	endolith
2	Spinach is high in calcium	by	conte
2	Spinach is a food edible by humans	by	Rosa
1	spinach is green	by	verbosity
1	spinach is green food	by	verbosity
1	some sandwiches contain spinach	by	qubyte
1	spinach is edible	by	openmind

Humans: Volunteers

- Wikidata: 18k active users
- Intrinsic motivation achieves great things
- Broad expertise, compared with selected experts or paid crowdsourcing
- https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all

Humans: Challenges

- ConceptNet:
  - Common knowledge, normalization
- Crowdsourcing: Quality assurance
- Wikidata: Modelling and agreement
  - E.g., ethnicity, notable_work, …
  - Multilingual concept alignment

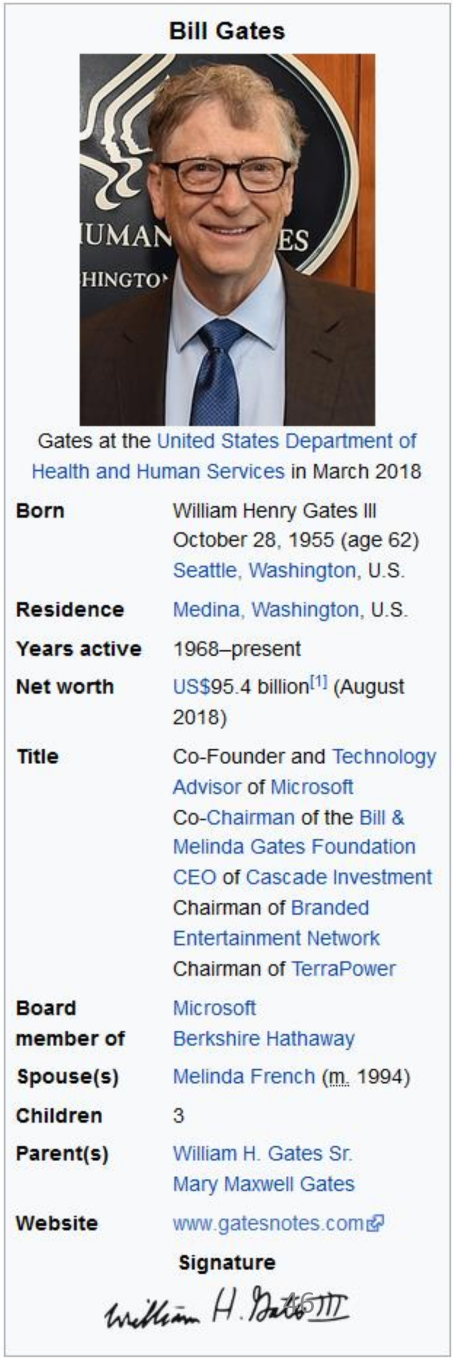


elephant is capable of...

- en carry a trunk →
- en forget to go on the paper →
- en lift logs from the ground →
- en to lift the tree →
- en remember water sources →
- en visit the grocery store →
- en weigh up to 14000 pounds →
- en weight 1000 kilos →

B. Structured extraction

- Wikipedia already provides structured data
- All we need to do is harvest…



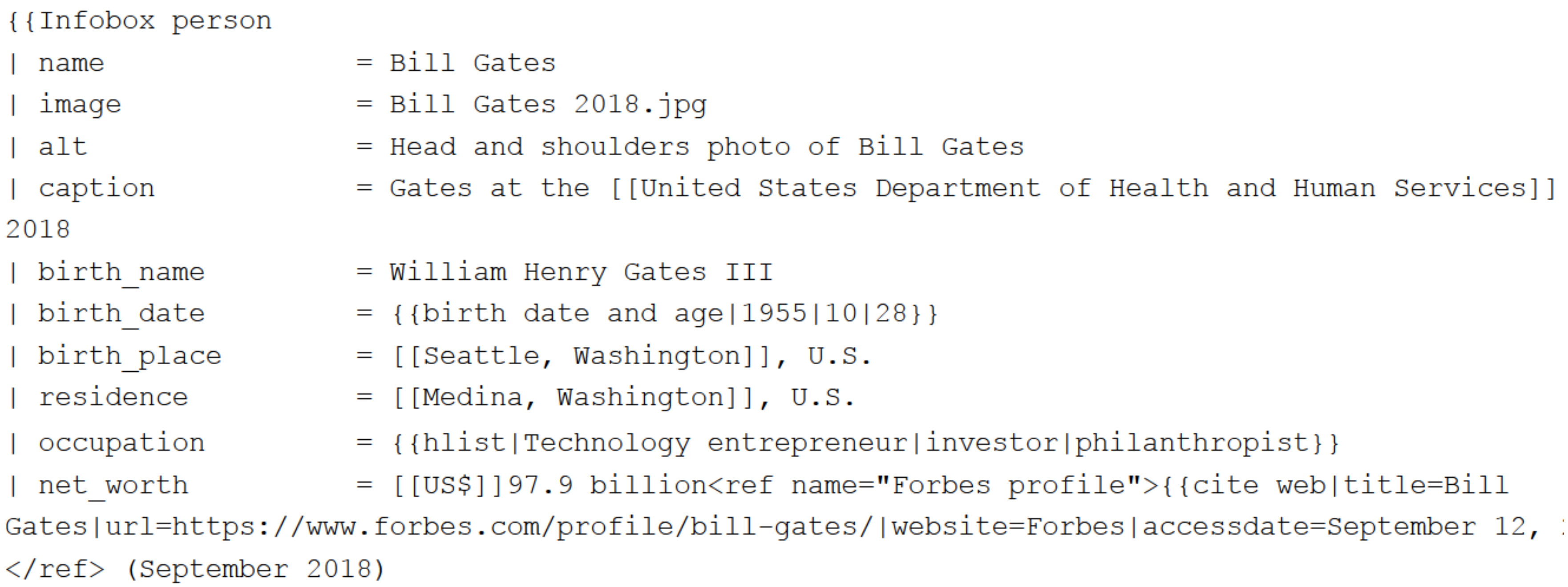
Bill Gates

Gates at the United States Department of Health and Human Services in March 2018

Born	William Henry Gates III October 28, 1955 (age 62) Seattle, Washington, U.S.
Residence	Medina, Washington, U.S.
Years active	1968–present
Net worth	US$95.4 billion[1] (August 2018)
Title	Co-Founder and Technology Advisor of Microsoft Co-Chairman of the Bill & Melinda Gates Foundation CEO of Cascade Investment Chairman of Branded Entertainment Network Chairman of TerraPower
Board member of	Microsoft Berkshire Hathaway
Spouse(s)	Melinda French (m. 1994)
Children	3
Parent(s)	William H. Gates Sr. Mary Maxwell Gates
Website	www.gatesnotes.com

Signature

William H. Gates III



```
{{Infobox person
| name                   = Bill Gates
| image                  = Bill Gates 2018.jpg
| alt                    = Head and shoulders photo of Bill Gates
| caption                = Gates at the [[United States Department of Health and Human Services]]
2018
| birth_name             = William Henry Gates III
| birth_date             = {{birth date and age|1955|10|28}}
| birth_place            = [[Seattle, Washington]], U.S.
| residence              = [[Medina, Washington]], U.S.
| occupation             = {{hlist|Technology entrepreneur|investor|philanthropist}}
| net_worth              = [[US$]]97.9 billion<ref name="Forbes profile">{{cite web|title=Bill
Gates|url=https://www.forbes.com/profile/bill-gates/|website=Forbes|accessdate=September 12,
</ref> (September 2018)
```

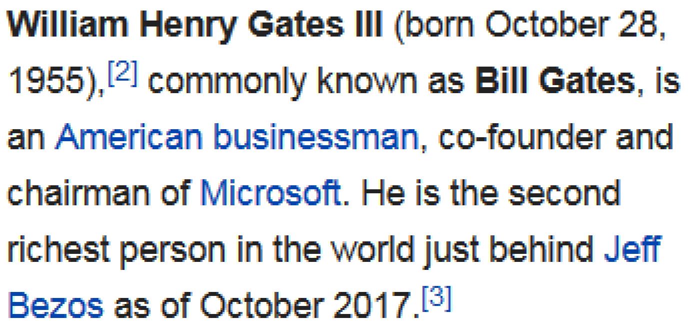
Work done?

- Noise
- Canonicalization of entities and predicates
- Usage of category system

Examples: YAGO, DBpedia

C. Text extraction

- In principle most powerful
  - No need for humans
  - No restriction to Wikipedia existence



William Henry Gates III (born October 28, 1955),[2] commonly known as **Bill Gates**, is an American businessman, co-founder and chairman of Microsoft. He is the second richest person in the world just behind Jeff Bezos as of October 2017.[3]

- In practice big noise challenges
  - Many pipeline steps
    - Named-entity recognition, named-entity disambiguation, relation extraction, relation canonicalization, extraction consolidation, ..

- Examples: NELL, Textrunner

Text extraction demo (relations part)

- https://www.rosette.com/capability/relationship-extraction/#try-the-demo

- *Merkel is of German and Polish descent. Her paternal grandfather, Ludwik Kasner, was a German policeman of Polish ethnicity, who had taken part in Poland's struggle for independence in the early 20th century.[22] He married Merkel's grandmother Margarethe, a German from Berlin, and relocated to her hometown where he worked in the police. In 1930, they Germanized the Polish name Kaźmierczak to Kasner.[23][24][25][26] Merkel's maternal grandparents were the Danzig politician Willi Jentzsch, and Gertrud Alma née Drange, a daughter of the city clerk of Elbing (now Elbląg, Poland) Emil Drange. Since the mid 1990s, Merkel has publicly mentioned her Polish heritage on several occasions and described herself as a quarter Polish, but her Polish roots became better known as a result of a 2013 biography.*
- *In 1968, Merkel joined the Free German Youth (FDJ), the official communist youth movement sponsored by the ruling Marxist–Leninist Socialist Unity Party of Germany.[30][31][32] Membership was nominally voluntary, but those who did not join found it difficult to gain admission to higher education.[33] She did not participate in the secular coming of age ceremony Jugendweihe, however, which was common in East Germany. Instead, she was confirmed.[34] During this time, she participated in several compulsory courses on Marxism-Leninism with her grades only being regarded as "sufficient".*

D. Constraints

Databases

- Key, foreign key, range, …

Knowledge bases:

- *Events start earlier than they end*
- *Every human must have two parents*
- *Mayors of cities must be humans*
- *The parent of a person’s sibling is the person’s parent*

- Can be used to…

  … reject KB modifications
  … indicate missing information
  … infer new facts

- But reality is messy..

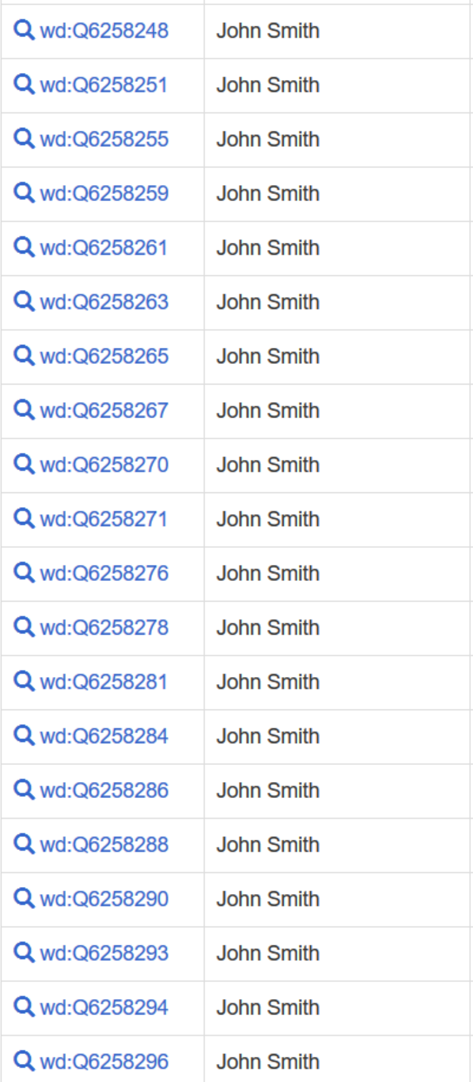
3. Introduction to AKBC

I. Motivation
II. Terminology
III. Topics
IV. Construction techniques
V. Applications
VI. Past, present and future

What KBs are good for

- Master data
- Data mining
- Search enhancements
- Question answering
- Language generation
- Entity linking
- Learning more knowledge
- ….

Master data (1)

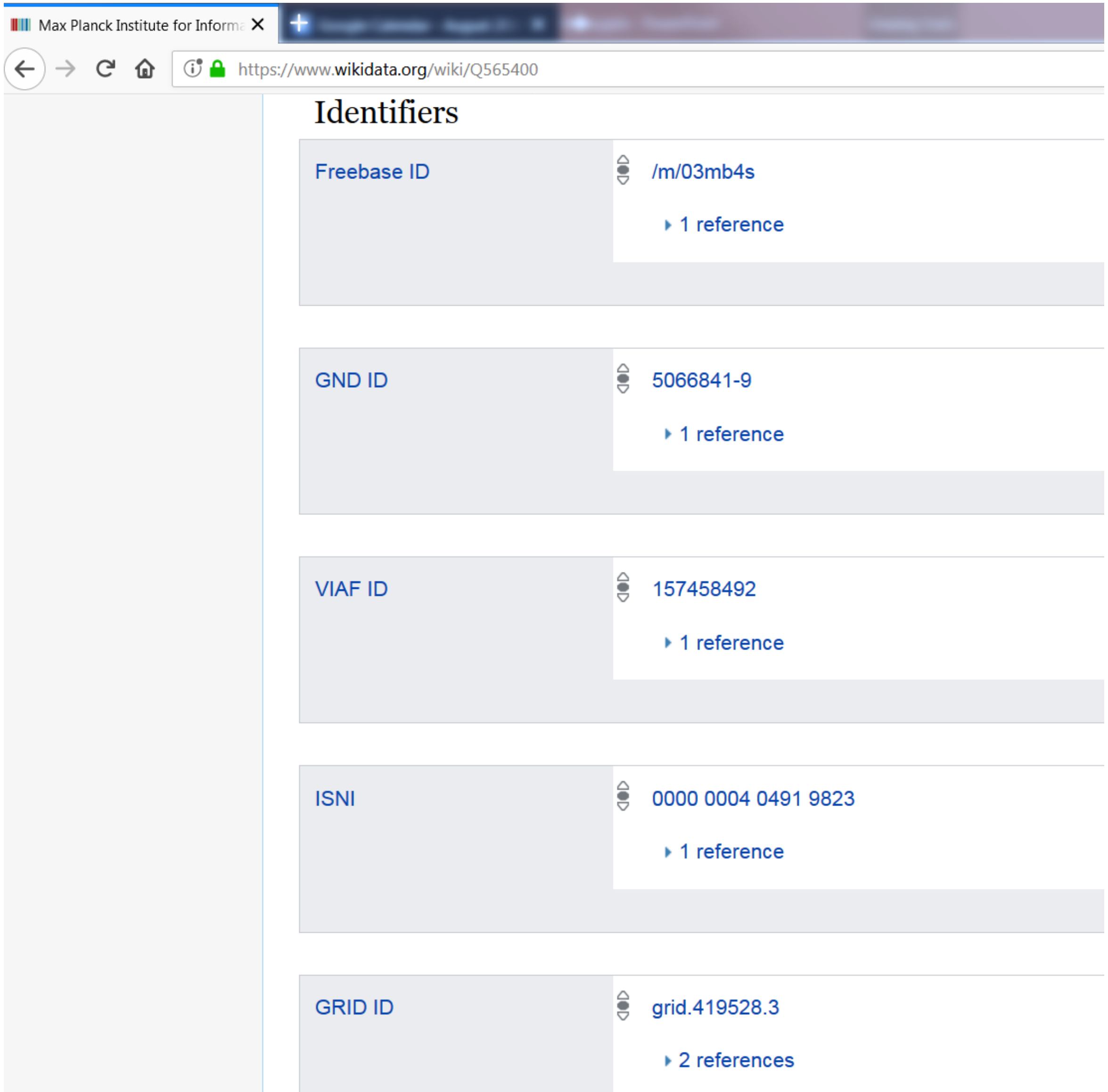


wd:Q6258248	John Smith
wd:Q6258251	John Smith
wd:Q6258255	John Smith
wd:Q6258259	John Smith
wd:Q6258261	John Smith
wd:Q6258263	John Smith
wd:Q6258265	John Smith
wd:Q6258267	John Smith
wd:Q6258270	John Smith
wd:Q6258271	John Smith
wd:Q6258276	John Smith
wd:Q6258278	John Smith
wd:Q6258281	John Smith
wd:Q6258284	John Smith
wd:Q6258286	John Smith
wd:Q6258288	John Smith
wd:Q6258290	John Smith
wd:Q6258293	John Smith
wd:Q6258294	John Smith
wd:Q6258296	John Smith

(300 more)

https://w.wiki/A4Z

Master data (2)



Max Planck Institute for Informa

https://www.wikidata.org/wiki/Q565400

Identifiers

Freebase ID	/m/03mb4s 1 reference
GND ID	5066841-9 1 reference
VIAF ID	157458492 1 reference
ISNI	0000 0004 0491 9823 1 reference
GRID ID	grid.419528.3 2 references

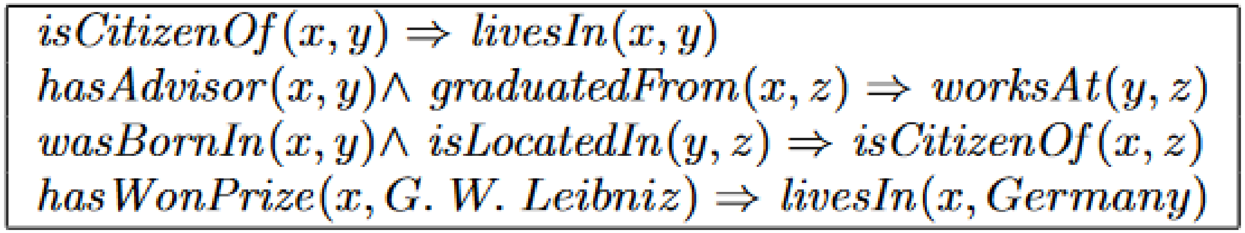
Relevant for:

- Museums
- Libraries
- Scientific publications

....

Data mining

- Use input facts to extract patterns that allow to predict new facts



$$isCitizenOf(x,y) \Rightarrow livesIn(x,y)$$
$$hasAdvisor(x,y) \wedge graduatedFrom(x,z) \Rightarrow worksAt(y,z)$$
$$wasBornIn(x,y) \wedge isLocatedIn(y,z) \Rightarrow isCitizenOf(x,z)$$
$$hasWonPrize(x, G.\ W.\ Leibniz) \Rightarrow livesIn(x, Germany)$$

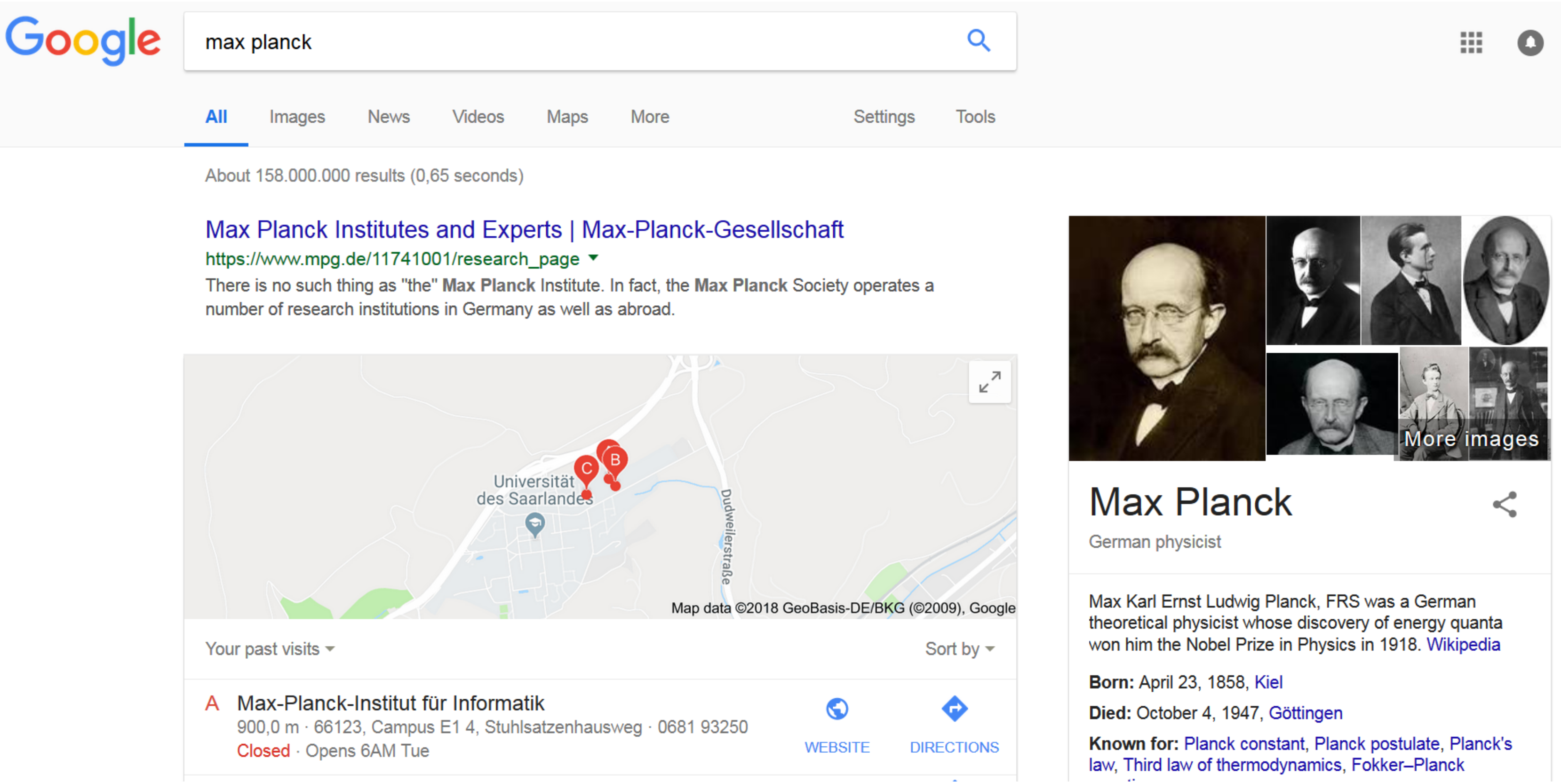
isCitizenOf(John, France) → livesIn(John, France)

- Various approaches based on association rule mining and latent models

Entity linking

https://opentapioca.org/

Search enhancements



About 158.000.000 results (0,65 seconds)

Max Planck Institutes and Experts | Max-Planck-Gesellschaft
https://www.mpg.de/11741001/research_page
There is no such thing as "the" **Max Planck** Institute. In fact, the **Max Planck** Society operates a number of research institutions in Germany as well as abroad.

Map data ©2018 GeoBasis-DE/BKG (©2009), Google

Your past visits ▾ Sort by ▾

A Max-Planck-Institut für Informatik
900,0 m · 66123, Campus E1 4, Stuhlsatzenhausweg · 0681 93250
Closed · Opens 6AM Tue

Max Planck
German physicist

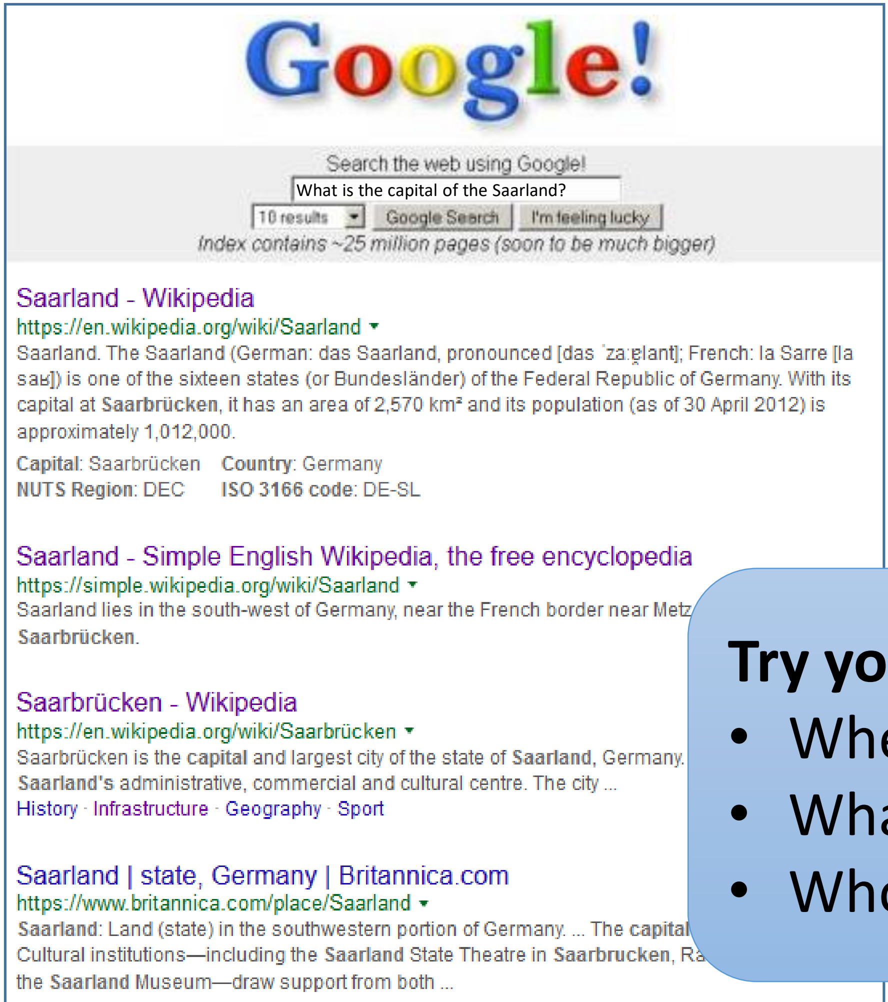
Max Karl Ernst Ludwig Planck, FRS was a German theoretical physicist whose discovery of energy quanta won him the Nobel Prize in Physics in 1918. Wikipedia

Born: April 23, 1858, Kiel

Died: October 4, 1947, Göttingen

Known for: Planck constant, Planck postulate, Planck's law, Third law of thermodynamics, Fokker–Planck

Question answering



Search the web using Google!

What is the capital of the Saarland?

10 results ▾ | Google Search | I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

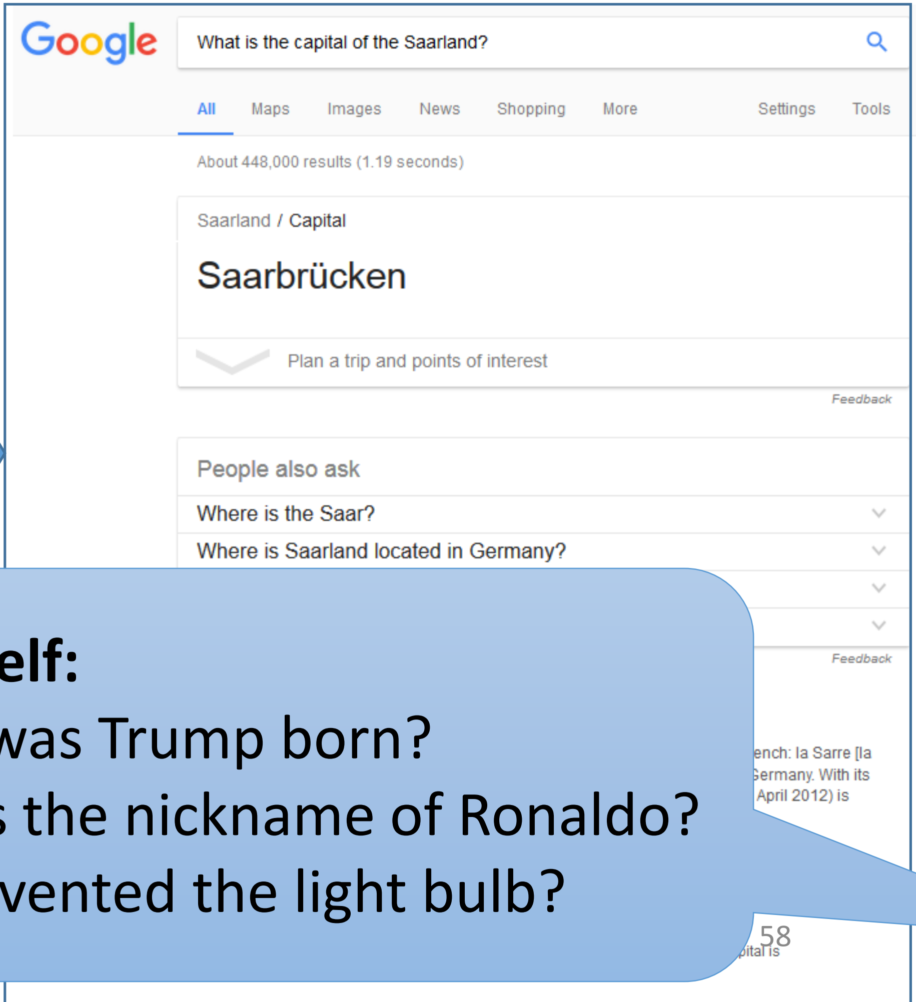
Saarland - Wikipedia
https://en.wikipedia.org/wiki/Saarland ▾
Saarland. The Saarland (German: das Saarland, pronounced [das ˈzaːɐ̯lant]; French: la Sarre [la saʁ]) is one of the sixteen states (or Bundesländer) of the Federal Republic of Germany. With its capital at **Saarbrücken**, it has an area of 2,570 km² and its population (as of 30 April 2012) is approximately 1,012,000.

Capital: Saarbrücken **Country**: Germany
NUTS Region: DEC **ISO 3166 code**: DE-SL

Saarland - Simple English Wikipedia, the free encyclopedia
https://simple.wikipedia.org/wiki/Saarland ▾
Saarland lies in the south-west of Germany, near the French border near Metz
Saarbrücken.

Saarbrücken - Wikipedia
https://en.wikipedia.org/wiki/Saarbrücken ▾
Saarbrücken is the **capital** and largest city of the state of **Saarland**, Germany.
Saarland's administrative, commercial and cultural centre. The city ...
History · Infrastructure · Geography · Sport

Saarland | state, Germany | Britannica.com
https://www.britannica.com/place/Saarland ▾
Saarland: Land (state) in the southwestern portion of Germany. ... The capital
Cultural institutions—including the **Saarland** State Theatre in **Saarbrucken**, Ra
the **Saarland** Museum—draw support from both ...



What is the capital of the Saarland?

About 448,000 results (1.19 seconds)

Saarland / Capital

Saarbrücken

Plan a trip and points of interest

Feedback

People also ask

Where is the Saar?

Where is Saarland located in Germany?

Feedback

ench: la Sarre [la
Germany. With its
April 2012) is

pital is

Try yourself:

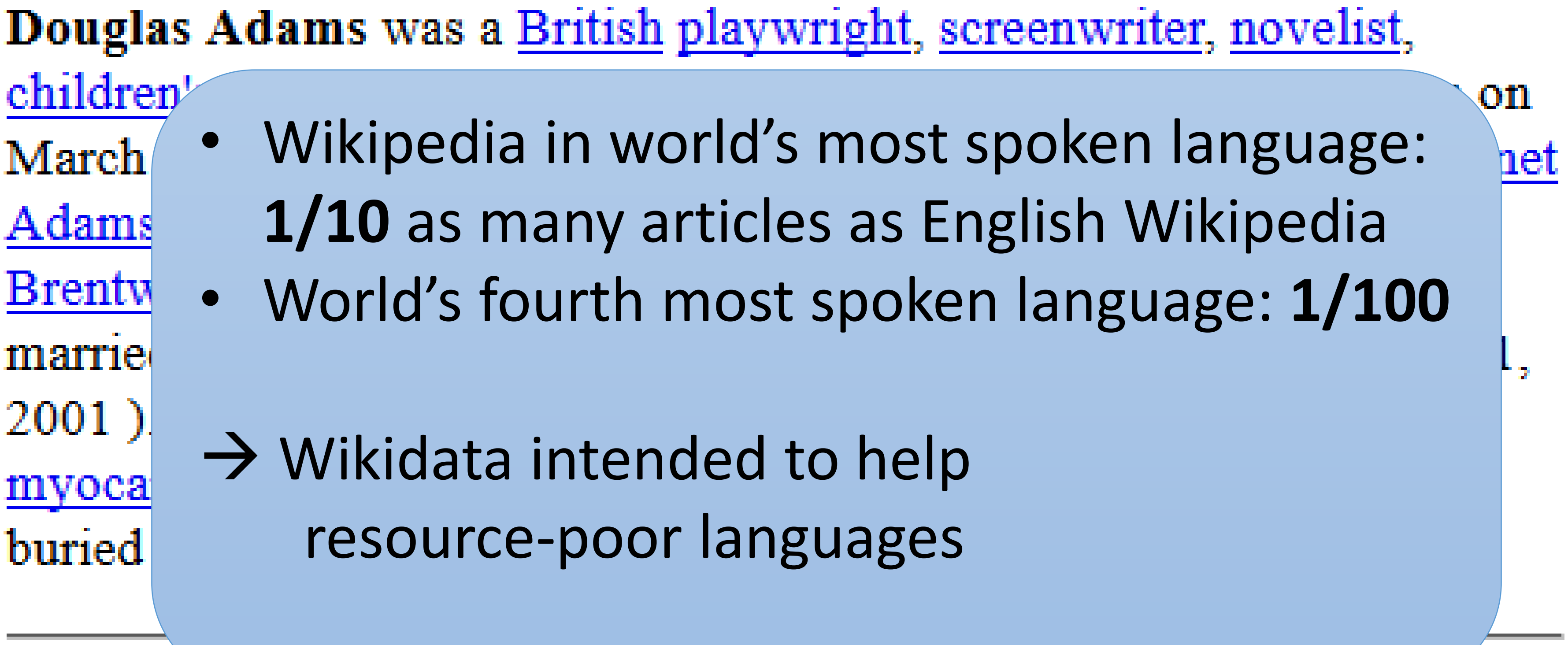
- When was Trump born?
- What is the nickname of Ronaldo?
- Who invented the light bulb?

Question answering (2)

- Knowledge bases key component in question answering systems
  - E.g., IBM Watson

- AllenAI science challenge: Computers currently in 8th grade
  - Knowledge acquisition still major bottleneck

Language generation



Douglas Adams was a British playwright, screenwriter, novelist,
children' … on
March … net
Adams
Brentw
marrie … l,
2001).
myoca
buried

- Wikipedia in world’s most spoken language: **1/10** as many articles as English Wikipedia
- World’s fourth most spoken language: **1/100**

→ Wikidata intended to help resource-poor languages

https://tools.wmflabs.org/autodesc?q=9021&lang=&mode=long&links=reasonator&redlinks=reasonator&format=html

3. Introduction to AKBC

I. Motivation

II. Terminology

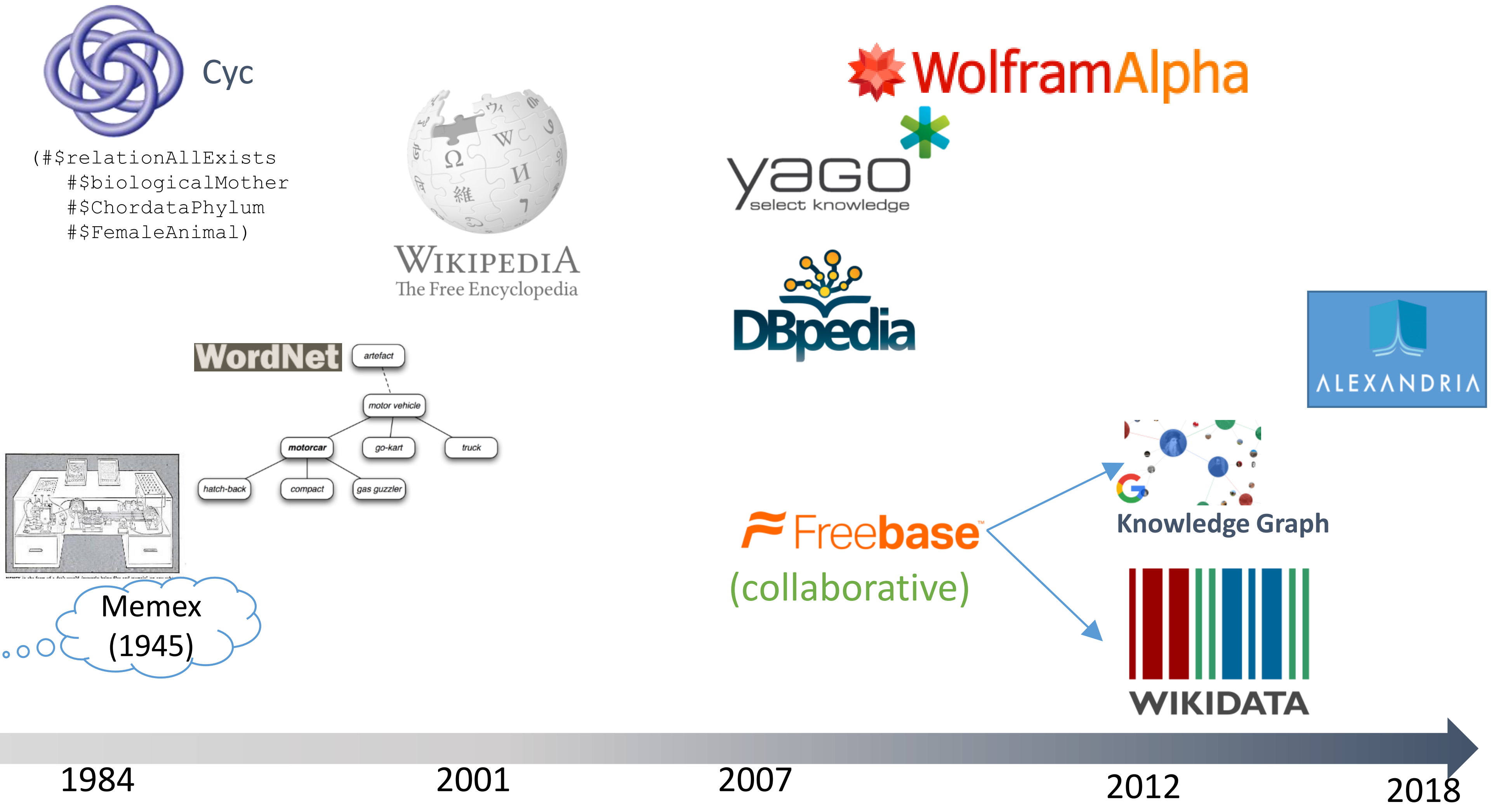
III. Topics

IV. Construction techniques

V. Applications

VI. Past, present and future

Past



```
(#$relationAllExists
   #$biologicalMother
   #$ChordataPhylum
   #$FemaleAnimal)
```

Present

- KBs at most major tech companies and beyond
  - Google, Microsoft, Alibaba, Bloomberg, …

- Feb 2018: $125 million investment by Microsoft cofounder Paul Allen into non-profit research on common sense knowledge extraction and reasoning

- Research: Major part of NLP conferences taken up by IE/AKBC research

Future

- ?

Outline

1. Introducing each other
2. Course organization
3. What, Why, How
4. **Lab 1**

Lab 1

- Information extraction where from?
    - Actual web crawling nontrivial
    - Wikipedia a popular high-quality resource
- Learn/practice text manipulation, perform some simple analyses, get to know KB querying

Take home

- Knowledge base construction builds machine-readable structured content from unstructured/semistructured inputs

- Structured data is relevant for a range of knowledge-intensive tasks

- Next week: Crawling and scraping